International Series in
Operations Research and Management Science

Richard J. Boucherie
Nico M. van Dijk *Editors*

Queueing Networks

A Fundamental Approach

Springer

International Series in Operations Research & Management Science

Volume 154

For further volumes:
http://www.springer.com/series/6161

Richard J. Boucherie • Nico M. van Dijk
Editors

Queueing Networks

A Fundamental Approach

Springer

Editors
Richard J. Boucherie
Departement of Applied Mathematics
University of Twente
Stochastic OR Group
PO Box 217
7500 AE Enschede
The Netherlands
r.j.boucherie@math.utwente.nl

Nico M. van Dijk
Faculty of Economics and Business
University of Amsterdam
Roetersstraat 11
1018 WB Amsterdam
The Netherlands
N.M.vanDijk@uva.nl

ISSN 0884-8289
ISBN 978-1-4419-6471-7 e-ISBN 978-1-4419-6472-4
DOI 10.1007/978-1-4419-6472-4
Springer New York Dordrecht Heidelberg London

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The origin of queueing theory and its application traces back to Erlang's historical work for telephony networks as recently celebrated by the Erlang Centennial, 100 Years of Queueing, Copenhagen, recalling his first paper in 1909. Ever since, the simplicity and fundamental flavour of Erlang's famous expressions, such as his loss formula for an incoming call in a circuit switched system to be lost, has remained intriguing. It has motivated the development of results with similar elegance and expression power for various systems modeling congestion and competition over resources.

A second milestone was the step of queueing theory into queueing networks as motivated by the first so-called product form results for assembly type networks in manufacturing in the nineteen fifties (R.R.P. Jackson 1954, J.R. Jackson 1957, and E. Koenigsberg 1958, 1959). These results revealed that the queue lengths at nodes of a network, where customers route among the nodes upon service completion in equilibrium can be regarded as independent random variables, that is, the equilibrium distribution of the network of nodes factorizes over (is a product of) the marginal equilibrium distributions of the individual nodes as if in isolation. These networks are nowadays referred to as Jackson networks.

A third milestone was inspired by the rapid development of computer systems and brought the attention for service disciplines such as the Processor Sharing discipline introduced by Kleinrock in 1967. More complicated multi server nodes and service disciplines such as First-Come-First-Served, Last-Come-First-Served and Processor Sharing, and their mixing within a network have led to a surge in theoretical developments and a wide applicability of queuing theory.

Queueing networks have obtained their place in both theory and practice. New technological developments such as Internet and wireless communications, but also advancements in existing applications such as manufacturing and production systems, public transportation, logistics, and health care, have triggered many theoretical and practical results.

Queueing network theory has focused on both the analysis of complex nodes, and the interaction between nodes in networks. This handbook aims to highlight fundamental, methodological and computational aspects of networks of queues to provide insight and unify results that can be applied in a more general manner. Several topics that are closely related are treated from the perspective of different authors to also provide different intuition that, in our opinion, is of fundamental importance to appreciate the results for networks of queues. Of course, applications of modern queueing networks are manifold. These are illustrated in the concluding chapters of this handbook. The handbook is organized in five parts.

Part 1. Exact analytical results, chapters 1–7

Product form expressions for the equilibrium distribution of networks are by far leading and have been most dominant in the literature on exact analytical results for queueing networks. In recent years, features such as batch routing, negative customers and signals have been introduced to enhance the modeling power of this class of networks. A unified theory from different perspectives is contained in the first part of this handbook. Topics include

- a characterization of product forms by physical balance concepts and simple traffic flow equations,
- classes of service and queue disciplines such as Invariant Disciplines and Order Independent queues that allow a product form,
- a unified description of product forms for discrete time queueing networks,
- insights for insensitivity from the classical Erlang loss model up to Generalised Semi Markov Processes and partially insensitive networks,
- aggregation and decomposition results that allow subnetworks to be aggregated into single nodes to reduce computational burden.

These product form results encompass a number of intriguing aspects that are not only most useful for practical purposes but also indicate a variety of open problems which remain to be tackled.

Part 2. Monotonicity and comparison results, chapters 8–9

Exact (product form) results are only available for a limited class of networks. These exact results, however, may also be invoked to obtain bounds for performance measures for intractable queueing networks. Two basic approaches can be identified:

- stochastic monotonicity and ordering results based on the ordering of the generators of the processes,
- comparison results and explicit error bounds based on an underlying Markov reward structure which leads to ordering of expectations of performance measures.

There is a clear trade-off for applying either of these two approaches. Stochastic monotonicity yields stronger results such as with non-exponential service times.

The Markov reward approach in turn is applicable under less stringent conditions, particularly with more complex structures as in a queueing network. These results are not only of theoretical and qualitative interest by themselves, but also motivate the derivation of exact analytical results to enable bounds.

Part 3. Diffusion and fluid results, chapters 10–12

Limiting regimes often allow for amenable expressions for performance measures in systems that are otherwise intractable. Two particular regimes are of interest: the fluid regime and the diffusion regime that are illustrated through the following topics:

- fluid limits for analysis of system stability,
- diffusion approximation for multi-server systems,
- system fed by Gaussian traffic to model variation in the arrival process.

These topics illustrate a rich class of systems that may be analyzed in the limiting regime and identify an important area of current research.

Part 4. Computational and approximate results, chapters 13–15

Practical applications such as in manufacturing, computer performance and communications rapidly prove to be beyond analytical solvability due to e.g. non-exponential service times, capacity constraints, synchronization or prioritization. Numerically exact or approximate approaches for averages or distributions of performance measures have been developed in literature. An illustration is provided via the following topics:

- MVA (mean value analysis) and QNA (queueing network analyzer) focusing on mean and variance of performance measures such as queue length and sojourn times,
- numerical approximation of response time distributions
- approximate decomposition results for large open queueing networks.

The numerical approach to performance analysis is a lively research community that considerably contributes to the success of queueing theory in applications as it allows for explicit numerical results for performance measures.

Part 5. Selected applications, chapters 16–18

Applications of queueing networks are manifold. To illustrate the application power of queueing theory, some special application areas and their specific queueing network aspects are enlightened:

- loss networks as originating from circuit switched telecommunications applications,
- capacity sharing as originating from packet switching in data networks,
- hospital logistics.

The first two applications have a theoretical nature as they illustrate a typical class of queueing networks. The last application illustrates a typical approach for application of queueing theory in a practical environment.

Despite the fundamental theoretical flavour of this book, it is to be kept in mind that the area of queueing theory would not have existed and would not have progressed so strongly had it not been driven by application areas that led to the various fundamental questions. The intertwined progress of theory and practice will remain to be most intriguing and will continue to be the basis of further developments in queueing theory. You are highly invited to step in.

Acknowledgments

We would like to express some words of gratitude. First of all, we are most grateful to all authors involved for their positive reactions and enthusiasm right from the start, for their cooperation and for their invaluable contributions. We wish to apologize to those authors who had already submitted at an early stage but who had to wait for such a long time until its final publication. We like to apologize to some other authors for our persistently hampering on their shoulders. We hope that you can appreciate our intention and final outcome.

We deeply owe our gratefulness to the editor-in-chief Dr. F.S. Hillier. Had it not been for his stimulation, patience and confidence in its finalization, we would not have been able to complete this book successfully.

The support by the Dutch Organisation for Scientific Research (NWO) by which the outcome of this book could be realized has been highly appreciated.

Enschede,
Amsterdam,
June 2010

Richard J. Boucherie
Nico M. Van Dijk

Contents

List of Contributors

Ivo Adan
Eindhoven University of Technology, the Netherlands,
and University of Amsterdam, the Netherlands
e-mail: i.j.b.f.adan@tue.nl

Varsha Apte
IIT Bombay, India
e-mail: varsha@cse.iitb.ac.in

Thomas Bonald
Orange Labs, France
e-mail: thomas.bonald@orange-ftgroup.com

Richard J. Boucherie
University of Twente, the Netherlands
e-mail: r.j.boucherie@utwente.nl

Xiuli Chao
University of Michigan, USA
e-mail: xchao@umich.edu

Hong Chen
University of British Columbia, Canada
e-mail: hong.chen@sauder.ubc.ca

Stefan Creemers
Catholic University of Leuven, Belgium
e-mail: stefan.creemers@econ.kuleuven.be

Hans Daduna
University of Hamburg, Germany
e-mail: daduna@math.uni-hamburg.de

J. G. Dai
Georgia Institute of Technology, USA
e-mail: dai@gatech.edu

Nico M. van Dijk
University of Amsterdam, the Netherlands
e-mail: n.m.vandijk@uva.nl

Michael Grottke
University of Erlangen-Nuremberg, Germany
e-mail: Michael.Grottke@wiso.uni-erlangen.de

John J. Hasenbein
University of Texas at Austin, USA
e-mail: jhas@mail.utexas.edu

Boudewijn R. Haverkort
University of Twente, the Netherlands,
and Embedded Systems Institute, the Netherlands
e-mail: boudewijn.haverkort@esi.nl

Tijs Huisman
ProRail, the Netherlands
e-mail: Tijs.Huisman@prorail.nl

Bara Kim
Korea University, Korea
e-mail: bara@korea.ac.kr

A.E. Krzesinski
University of Stellenbosch, South Africa
e-mail: aek1@cs.sun.ac.za

Marc Lambrecht
Catholic University of Leuven, Belgium
e-mail: marc.lambrecht@econ.kuleuven.be

Michel Mandjes
University of Amsterdam, the Netherlands
e-mail: m.r.h.mandjes@uva.nl

Masakiyo Miyazawa
Tokyo University of Science, Japan
e-mail: miyazawa@is.noda.tus.ac.jp

Alexandre Proutière
Microsoft Research, UK
e-mail: alexandre.proutiere@microsoft.com

Ramin Sadre
University of Twente, the Netherlands
e-mail: r.sadre@utwente.nl

Ryszard Szekli
University of Wrocław, Poland
e-mail: Ryszard.Szekli@math.uni.wroc.pl

P.G. Taylor
University of Melbourne, Australia
e-mail: p.taylor@ms.unimelb.edu.au

Kishor S. Trivedi
Duke University, USA
e-mail: kst@ee.duke.edu

Jan van der Wal
Eindhoven University of Technology, the Netherlands,
and University of Amsterdam, the Netherlands
e-mail: jan.v.d.wal@tue.nl

Steve Woolet
IBM Corporation, USA
e-mail: steve_woolet@us.ibm.com

Heng-Qing Ye
Hong Kong Polytechnic University, China
e-mail: lgtyehq@inet.polyu.edu.hk

Stan Zachary
Heriot-Watt University, UK
e-mail: s.zachary@hw.ac.uk

Ilze Ziedins
University of Auckland, New Zealand
e-mail: i.ziedins@auckland.ac.nz

Chapter 1
On Practical Product Form Characterizations

Nico M. van Dijk

Abstract

Do we have a product form?
If so, how is it characterized?
If not, how can product forms still be useful?

The first question is not be that easy as it seems. The answer might depend on the state level of interest, the service assumptions imposed and the system restrictions or flexibilities in order. This chapter aims to address these question in two parts:

A: Product Forms: A Single Station
B: Product Forms: Tandem and Cluster Structures

In **A** just a single service station is studied to show how different levels of a state description and notions of balance may lead to analytic forms that can be referred to as 'product forms'. It covers simple birth-death type systems, forms of access blocking for multi-class stations, and symmetric up to so-called invariant disciplines.

In **B** just a tandem type structure (that is with consecutive service stations) and some Jacksonian cluster extensions are dealt with to show:

(i) The effect on the existence of a product form under practical phenomena as blocking and service sharing

(ii) How this existence can be characterized
- in an analytic manner by 'adjoint' reversibility
- by simple physical station or cluster 'outrate=inrate' principles

(iii) The practical way in which these insights can be used to obtain simple product form bounds for únsolvable systems

Nico M. van Dijk
University of Amsterdam, Roetersstraat 11, 1018 WB Amsterdam, The Netherlands
e-mail: n.m.vandijk@uva.nl

R.J. Boucherie and N.M. van Dijk (eds.), *Queueing Networks: A Fundamental Approach*,
International Series in Operations Research & Management Science 154,
DOI 10.1007/978-1-4419-6472-4_1, © Springer Science+Business Media, LLC 2011

A: Product Forms: Single Station Hierarchy

1.1 Introduction

No doubt that the popularity of queueing networks, next to its potential for modeling a variety of practical service networks, is largely due to the existence of the well-known product form expressions ever since the pioneering work by Jackson (1957).

In several subsequent chapters, various generalizations of this product form in more abstract settings will be provided (and different lights will be shed on its validity). For further interest in these generalizations the reader is referred to these chapters.

Nevertheless, as of today, a number of questions related to the existence of product forms still seem to be open, most notably among which:

- The simple question whether a specific network of interest has a product form or not.
- What is actually meant by a product form, is it uniquely defined, and under which conditions and to what extent or level does it apply.
- Last but not least, how can we guess and verify a particular product form in a down-to-earth manner.

This chapter merely aims to provide some more insights and partial answers for these questions. It also aims to do so in an instructive manner by following the down-to-earth approach of straightforward verification of balance equations. As such, it will be far from exhaustive. Roughly the objectives are:

Objectives.

1. To show the verification and the relation of product forms with different (levels of) partial balance and to emphasize the physical interpretation of these partial balances.
2. To show (a hierarchy of) different levels of product forms and partial balance as depending on state description and conditions satisfied.
3. To show the characterization of these partial balances and its related product forms by means of reversibility and, as will be called:
   'adjoint reversibility'.
4. To provide instructive as well as 'non-standard' product form examples, some of which might still be regarded as 'new', or at least which have not been reported explicitly.
5. To illustrate the application of these product form insights such as to provide simple and possibly insensitive product form bounds for practical non-product form systems.

Outline and results.

In line with the first three objectives, in **A** (sections 1.2-1.4) just a single service station will be analyzed. More precisely, three levels of balance will be considered

- for a total number at a station
- for each job-class separately
- for each service position

First in section 1.2, it is shown that these three levels lead to the well-known categories of

(i) birth-death systems
(ii) coordinate convex access structures and
(iii) insensitive symmetric service disciplines

It is shown that these three categories and their hierarchy can be generalized to 'product form' structures for single service stations with more general access/blocking mechanisms and service disciplines. Among others these results cover and extend classical results for multiclass service stations. Next, in section 1.3 the symmetric disciplines and its insensitivity are generalized to service invariant disciplines. These include some 'nonstandard' examples. Section 1.4 completes **A** with a direct application of simple bounds for $M|G|c|c+m$ systems, a literature discussion and an overview of the balances.

Next, in **B** (sections 1.5-1.7) tandem type networks are dealt with. First, in section 1.5, the most simple and generic but nón-reversible 'network structure' of a simple but finite two station open tandem queue is extensively studied. Despite its nón-reversibility it is shown how product forms can be concluded as based upon an artificially constructed adjoint tandem queue and an extended form of reversibility, which will be called 'adjoint reversibility'. This characterization by 'adjoint reversibility' leads to a sufficient and necessary characterization of a product form.

Various product form examples can so still be concluded also for tandem queues with finite capacities (blocking) or service sharing (fair sharing), such as of practical interest for manufacturing or internet modeling.

In section 1.6 extensions are provided with a single service station replaced by a Jackson network.

In section 1.7 it is argued and also numerically illustrated how these product form results may provide useful bounds for more natural but únsolvable systems with blocking. A simple optimal design application is included.

Section 1.8 shows how the results from **B** led to a special recent practical application for hospitals

Section 1.9 completes **B** with a literature discussion, a brief review and some remaining open questions for research.

1.2 Product Forms: Three Balances

Product forms are generally associated with a closed form expression that factorizes into separate terms for separate service stations. But in fact, such factorizations also exist for just one service station as by characterizations of partial balance. To shed some more light on this phenomenon of a product form characterization, in **B** we will simply consider a single service station.

1.2.1 Station Balance: B-D or Erlang-Engset systems

Consider a single server station. With n the number of jobs present, let

$$\begin{cases} \lambda(n) & \text{: be the arrival rate} \\ \mu(n) & \text{: be the service rate} \end{cases}$$

where it is assumed that $\mu(n) > 0$ for $n > 0$ and where $\lambda(n) > 0$ for all $n < N$, with N some finite number or infinite. (More precisely, that is, with n jobs present the arrival time up to the next arrival is exponentially distributed with parameter $\lambda(n)$ and similarly the time up to a next departure with parameter $\mu(n)$).

Note that no further specification is given on possible different job-types or on a possible service discipline, (e.g. a first-come first-served or other discipline in the single server case with $\mu(n) = \mu$ for all $n > 0$).

Let $\{\pi(n)\}$ represent the steady state distribution for the number of jobs present. This distribution is uniquely determined (up to its normalization) by the global balance (backwards Chapman-Kolmogorov) equations:

$$\left\{ \begin{matrix} \pi(n)\lambda(n) + \\ \pi(n)\mu(n) \end{matrix} \right\} = \left\{ \begin{matrix} \pi(n+1)\mu(n+1) + \\ \pi(n-1)\lambda(n-1) \end{matrix} \right\} \qquad \begin{matrix} (1.1.1) \\ (1.1.2) \end{matrix} \qquad (1.1)$$

for any $n \leq N$. These equations can be interpreted as "out = in"-stream equations in the "mathematical" sense of bringing you out of (left hand side) and bringing you into (right hand side) state n. However, substituting $n = 0$, necessarily requires that

$$\begin{cases} \pi(0)\lambda(0) = \pi(1)\mu(1) \\ \Longleftrightarrow (2.1.1) \text{ for } n = 0 \Longleftrightarrow (2.1.2) \text{ for } n = 1 \\ \Longrightarrow \ \pi(1)\lambda(1) = \pi(2)\mu(2) \Longleftrightarrow (2.1.1) \text{ for } n = 1 \Longleftrightarrow (2.1.2) \text{ for } n = 2 \end{cases}$$

and so on. Hence, the global balance relation (1.1) necessarily requires that

$$\pi(n)\mu(n) = \pi(n-1)\lambda(n-1) \Longleftrightarrow (2.1.1) \text{ for } n \geq 0 \Longleftrightarrow (2.1.2) \text{ for } n > 0. \quad (1.2)$$

Clearly, the relations (1.2) in turn, which are also well-known as 'birth-death equations', are sufficient for (1.1) to be satisfied. Though these relations are completely standard, in contrast with the 'mathematical' out = in interpretation of (1.1), an important point that might be emphasized here, is that the relations (1.2) have the more detailed interpretation that for any state n:

$$\begin{aligned}&\text{The (physical) outrate due to a departure} =\\&\text{the (physical) inrate due to an arrival at the service station.}\end{aligned}\qquad(1.3)$$

Here the outrate and inrate are to be read in the mathematical sense of leaving and entering that state respectively, as in (1.1). In this chapter this form of (physical) balance for a service station will be referred to as station balance (SB), so as to distinguish from more detailed notions of balances that will follow.

For $N < \infty$, and for $N = \infty$ under the standard ergodicity assumption that a probability solution $\pi(n)$ of (1.2) exists, roughly speaking that is, that a normalization constant c can be computed, (1.2) is satisfied by

$$\pi(n) = c\prod_{k=0}^{n-1}\left[\frac{\lambda(k)}{\mu(k+1)}\right] \qquad 0 \le n \le N \qquad (1.4)$$

Example 1.2.1 (Erlang systems) *As a most standard example, for given values s and m and by setting*

$$\lambda(n) = \lambda 1_{(n<s+m)}$$
$$\mu(n) = \begin{cases} n\mu & (n < s)\\ s\mu & (n \ge s)\end{cases}$$

the standard $M|M|s|s+m$ is included, with s servers and a waiting room of size m (possibly $m = \infty$), also known as an Erlang-system. (Erlang's pure delay system by $m = \infty$; Erlang's pure loss system by $m = 0$). In fact, even for these standard multiserver systems, the form (1.4) can already be regarded as a product form, as will be made more explicit by the following example.

Example 1.2.2 (Engset or Machine-Repair systems) *By*

$$\lambda(n) = (M-n)\gamma 1_{(n<s+m)} \qquad (n < M)$$

where M is some given finite (integer) number, and $\mu(n)$ as in example 1, we can also incorporate a finite source system with M sources (e.g. machines) each of which independently generates a service request (e.g. for a repair or maintenance) at an exponential rate γ. The service systems can be seen as in example 1 with s servers and a waiting facility of size m, hence a finite capacity for at most $N = s+m$ jobs (with $N < M$). When N sources (machines) are already in service mode, a next request is cancelled and a new request by that source is to be (exponentially) regenerated at rate γ.

As opposed to an Erlang system, in teletraffic theory this system is known as an Engset system. More generally it is also referred to as a Machine-Repair system. As a special case, for $s = N$, (1.4) reduces to

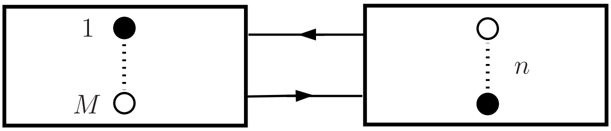


Fig. 1.1: Engset (Machine Repair) system.

$$\pi(n) = \tilde{c}\binom{M}{n}\left(\frac{1}{\mu}\right)^{n}\left(\frac{1}{\gamma}\right)^{M-n} \qquad \textit{with } \tilde{c} = c\gamma^{M}$$

For $M \leq N$ this is intuitively obvious as if each source can be seen as having its own devoted server and alternates between an operative mode, for an average length of time $(1/\gamma)$ and a service mode, for an average length of time $(1/\mu)$. For $N < M$, however, this first intuition seems less justified as a source may have multiple repeated operative sessions when all servers are busy. Nevertheless the form still applies.

A Product Form. More precisely, with M fixed, $\boldsymbol{n} = (n_1, n_2)$ where $n_1 = M - n$ and $n_2 = n$, $\gamma(k) = \lambda(M-k)$ we can also rewrite (1.4) as:

$$\pi(n) = \pi(n_1, n_2) = \tilde{\tilde{c}}\left[\prod_{k=1}^{n_1}\gamma(k)\right]^{-1}\left[\prod_{k=1}^{n_2}\mu(k)\right]^{-1} \tag{1.5}$$

with

$$\tilde{\tilde{c}} = c\left[\prod_{k=1}^{M}\gamma(k)\right],$$

This form can be regarded as a first representation of a product form in that it factorizes in stations: in this case: a *source* station and a *service* station, with a service rate $\gamma(k)$ and $\mu(k)$ respectively, when k jobs are present at that station.

1.2.2 Class balance: Coordinate convex property (CCP)

1.2.2.1 Two class coordinate convex case

Now consider a service station in which we can distinguish different job-classes. For its instructive purpose first assume two job-classes which can be regarded as independent up to common capacity constraints, say as determined by some set of admissible states $\boldsymbol{C} = \{(m_1, m_2) \mid m_1 \geq 0, m_2 \geq 0\}$. More precisely, with state

$\boldsymbol{m} = (m_1, m_2)$ denoting by

m_i : the numbers of jobs of type i present, for $i = 1,2$.

let

$$\begin{cases} \lambda_1(\boldsymbol{m}) &= \lambda_1 1_{(m_1+1,m_2)\in \boldsymbol{C}} \text{ be the arrival rate for job type 1} \\ \lambda_2(\boldsymbol{m}) &= \lambda_2 1_{(m_1,m_2+1)\in \boldsymbol{C}} \text{ the arrival rate for job type 2, and} \\ \mu_i(\boldsymbol{m}) &= \mu_i(m_i) \qquad\qquad \text{be the service rate for job type } i = 1,2. \end{cases}$$

Then, as in section 1.2.1, the steady state distribution $\{\pi(\boldsymbol{m})\}$ is uniquely determined (up to normalization) by the global balance equation which require that for any $\boldsymbol{m} \in \boldsymbol{C}$:

$$\begin{array}{ll}
\left\{ \begin{array}{l} \pi(m_1,m_2)\lambda_1 1_{((m_1+1,m_2)\in \boldsymbol{C})}+ \\ \pi(m_1,m_2)\lambda_2 1_{((m_1,m_2+1)\in \boldsymbol{C})}+ \\ \pi(m_1,m_2)\mu_1(m_1)+ \\ \pi(m_1,m_2)\mu_2(m_2) \end{array} \right\} & \begin{array}{l} (1.6.1) \\ (1.6.2) \\ (1.6.3) \\ (1.6.4) \end{array} \\
= & \\
\left\{ \begin{array}{l} \pi(m_1+1,m_2)\mu_1(m_1+1)1_{((m_1+1,m_2)\in \boldsymbol{C})}+ \\ \pi(m_1,m_2+1)\mu_2(m_2+1)1_{((m_1,m_2+1)\in \boldsymbol{C})}+ \\ \pi(m_1-1,m_2)\lambda_1+ \\ \pi(m_1,m_2-1)\lambda_2 \end{array} \right\} & \begin{array}{l} (1.6.1)' \\ (1.6.2)' \\ (1.6.3)' \\ (1.6.4)' \end{array}
\end{array} \tag{1.6}$$

In general, a simple analytic solution will no longer be available unless the more detailed relations $(1.6.i) = (1.6.i)'$ can be verified for $i = 1,\ldots,4$. For example, with the natural assumptions of just a common constraint for M jobs, i.e.

$$1_{\boldsymbol{C}}(m_1+1,m_2) = 1_{\boldsymbol{C}}(m_1,m_2+1) = 1_{(m_1+m_2+1\leq M)}$$

one directly verifies these detailed equations by

$$\pi(m_1,m_2) = c\left[\prod_{k=1}^{m_1}\frac{\lambda_1}{\mu_1(k)}\right]\left[\prod_{k=1}^{m_2}\frac{\lambda_2}{\mu_2(k)}\right] \tag{1.7}$$

Coordinate Convex Property (CCP). More generally, expression (1.7) satisfies $(2.6.i) = (2.6.i)'$, for $i = 1,\ldots,4$, provided the set $\boldsymbol{C}$ is coordinate convex, i.e.

$$(m_1,m_2) \in \boldsymbol{C} \Longrightarrow \begin{cases} (m_1-1,m_2) \in \boldsymbol{C} & (m_1 > 0) \\ (m_1,m_2-1) \in \boldsymbol{C} & (m_2 > 0) \end{cases}$$

Roughly speaking that is, $\boldsymbol{C}$ may not contain 'holes'. This condition is quite natural, such as in telecommunication structures, as will be illustrated below by some examples.

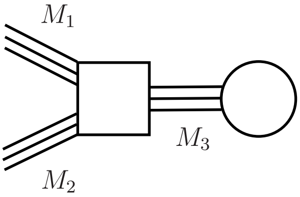


Fig. 1.2: End-to-end finite link groups.

Example 1.2.3 (Circuit switch) *As a simple circuit switch communication example consider type 1 and type 2 calls each of which with its own limited trunk group of M_1 and M_2 local channels connected to a common, say regional, trunk group of M channels. A call simultaneously requires, a local and regional channel during its entire call duration. With m_i the number of ongoing type-i calls, **C** is coordinate convex by*

$$\boldsymbol{C} = \{(m_1, m_2) \mid 0 \le m_1 \le M_1 \; ; \; 0 \le m_2 \le M_2 \; ; \; m_1 + m_2 \le M\}$$

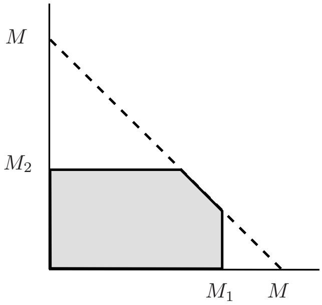


Fig. 1.3: CCP for circuit switch.

Example 1.2.4 (Overflow with call packing) *Now consider the most simple situation with type-1 calls at some finite primary trunk group with N_1 channels and type -2 calls at some second trunk group with N_2 channels. If all N_1 channels are busy a type-1 call can be overflowed to a free channel of the second group. Type-2 calls can only be handled by the second group. If an incoming type-1 or type-2 call cannot find an available channel as specified, it is rejected and lost. In addition (also see section 1.2.2.3), the so-called principle of call packing (or repacking) is assumed. That is, when a type-1 call at the primary group is completed, the channel that has become available takes over a type-1 call form the second group, if any. In this present setting this principle may seem unrealistic. Nevertheless, we refer to section 1.2.2.3 for further explanation of its 'necessity' as well as also a possible practical motivation.*

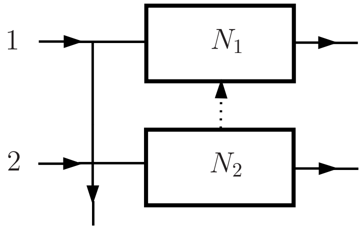


Fig. 1.4: Overflow system with CP.

Now note that under the call packing principle it suffices to keep track of just the total numbers m_1 and m_2 of ongoing type-1 and type-2 calls as specified by the coordinate convex region

$$\boldsymbol{C} = \{(m_1, m_2) \mid m_2 + (m_1 - N_1)^+ \leq N_2 \; ; \; m_1 \geq 0 \; ; \; m_2 \geq 0\}$$

As a consequence, under the the call packing principle, the form (1.7) thus applies for this overflow problem.

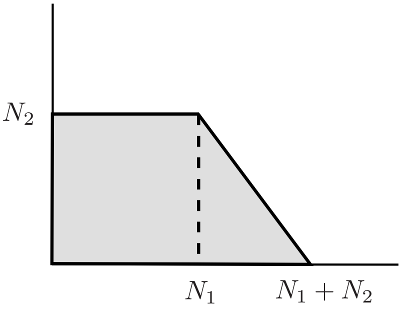


Fig. 1.5: CCP for overflow with CP.

Example 1.2.5 (Specialized servers) *As a slightly more generalized but possibly also more natural extension of the overflow situation in example 1.2.4, now consider two type of service requests, e.g. by critical and regular patients for intensive and medium care beds in a hospital, each with its own devoted group of S_1 and S_2 servers, e.g. special beds with associated nurses and equipment. In addition, however, if all type-1 servers are busy also type-2 servers, up to a maximum of K, can be used for type 1 requests, as far as available. The service times are request dependent. If no server is available a service request is rejected (e.g. in hospital practice some transfer takes place).*

Again, also a call packing principle is to be assumed to provide the product form solution (1.7). In this case of specialized servers (e.g. beds), however, it may even be *natural. More precisely, when a type-1 service (e.g. at an ICU bed) at a type 1 server is completed, a type-1 service (patient) at a (lower preference) type-2 server (e.g. a medium care bed) is 'switched' to this (higher preference) type-1 server (again see section 1.2.2.3).*

With m_i the number of ongoing type i servers, $\mathbf{C}$ is coordinate convex as shown in figure 1.6, as specified with $K \leq S_2$ by:

$$\begin{cases} m_1 \leq S_1 + K \\ m_2 + (m_1 - S_1)^+ \leq S_2 \end{cases}$$

which for $K = S_2$ reduces to:

$$m_2 + (m_1 - S_1)^+ \leq K = S_2$$

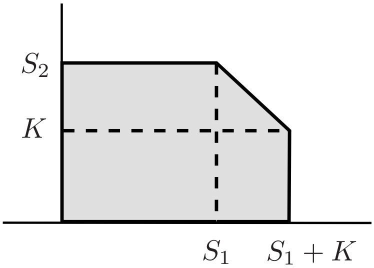


Fig. 1.6: CCP for specialized servers.

Remark 1.2.6 *Note that the practical descriptions and mechanisms in example 1.2.4 and 1.2.5 are quite different but that the product form solutions and the graphical representations of the admissible regions $\mathbf{C}$ are identical in form.*

1.2.2.2 Class Balance and Product Form

The two class situation can directly be generalized to R job classes. Let $\boldsymbol{m} = (m_1, m_2, \ldots, m_R)$ denote by m_r the number of jobs of job class r present, $r = 1, \ldots, R$.

Let e_r denote the r-th unit vector with the r-th component equal to 1 and 0 otherwise. Hence $\boldsymbol{m} + e_r = (m_1, \ldots, m_{r-1}, m_r + 1, m_{r+1}, \ldots, m_R)$ and similarly for $\boldsymbol{m} - e_r$. With

$$\begin{cases} \lambda_r(\boldsymbol{m}) = \lambda_r 1_{(\boldsymbol{m}+e_r \in \mathbf{C})} & \text{the arrival rate and} \\ \mu_r(\boldsymbol{m}) = \mu_r(m_r) & \text{the service rate} \end{cases}$$

for $r = 1,\dots,R$ and $\boldsymbol{C}$ some set of admissible states, the global balance equation then requires that for any state $\boldsymbol{m} \in \boldsymbol{C}$:

$$\begin{array}{lll} \left\{\begin{array}{l} \pi(\boldsymbol{m})\sum_r \mu_r(m_r)+ \\ \pi(\boldsymbol{m})\sum_r \lambda_r 1_{(\boldsymbol{m}+e_r\in\boldsymbol{C})} \end{array}\right\} & \begin{array}{l}(1.8.1)\\(1.8.2)\end{array} & \\ = & & (1.8) \\ \left\{\begin{array}{l} \sum_r \pi(\boldsymbol{m}-e_r)1_{(\boldsymbol{m}-e_r\in\boldsymbol{C})}\lambda_r+ \\ \sum_r \pi(\boldsymbol{m}+e_r)1_{(\boldsymbol{m}+e_r\in\boldsymbol{C})}\mu_r(m_r+1) \end{array}\right\} & \begin{array}{l}(1.8.1)'\\(1.8.2)'\end{array} & \end{array}$$

These are directly verified by each job class r separately by $(1.8.i) = (1.8.i)'$ for $i = 1,2$ and with $m_r > 0$:

$$\pi(\boldsymbol{m})\mu_r(m_r)1_{(\boldsymbol{m}\in\boldsymbol{C})} = \pi(\boldsymbol{m}-e_r)\lambda_r 1_{(\boldsymbol{m}-e_r\in\boldsymbol{C})} \tag{1.9}$$

provided $\boldsymbol{C}$ is coordinate convex, that is:

$$\boldsymbol{m}\in\boldsymbol{C} \Longrightarrow \boldsymbol{m}-e_r\in\boldsymbol{C} \qquad (\text{if } m_r > 0) \tag{1.10}$$

Relation (1.9) directly leads to the solution at $\boldsymbol{C}$:

$$\pi(\boldsymbol{m}) = c\prod_r \left[\prod_{k=1}^{m_r} \frac{\lambda_r}{\mu_r(k)}\right] \tag{1.11}$$

This steady state expression (1.11) can be referred to as a ’product form’ in that it factorizes to the steady state solutions for each job class separately with arrival rate λ_r and service rates $\mu_r(m_r)$ as if these are completely independent up to a common admissability region $\boldsymbol{C}$.

Furthermore, one may note that (1.9) states that for any job class r:

$$\begin{array}{l}\textit{The rate out of any state due to a class r departure } = \\ \textit{the rate into this state due to a class r arrival}\end{array} \tag{1.12}$$

In the present setting of this chapter (1.12) will be referred to as *class balance* (CB). Again a detailed and physical notion of an outrate = inrate interpretation (in this case by (1.12)) thus seems to be directly related to a ’product form’ solution in that it factorizes to the individual components of this detailed balance.

Remark 1.2.7 (Population distribution) *Clearly, by a product form expression as in (1.7) or (1.11) we can also, compute the steady state distribution $\pi(n)$ for the total number of jobs present $n = m_1 + m_2$ by*

$$\pi(n) = c\sum\nolimits_{\{(m_1,m_2)|m_1+m_2=n\}} \pi(m_1,m_2) \tag{1.13}$$

For example, in example 1.2.3 with $M_1 = M_2 = \infty$ and just a common capacity constraint $m_1 + m_2 \le M$, for the pure multiserver case with $\mu_i(m_i) = m_i\mu_i$, one *directly obtains*

$$\pi(n) = c\sum\nolimits_{(m_1,m_2)}\left[\prod_{i=1}^{2}\frac{1}{m_i!}\left(\frac{\lambda_i}{\mu_i}\right)^{m_i}\right] = c\frac{1}{n!}(\lambda\tau)^n\,,$$

$$\tau = \left[\frac{\lambda_1}{\lambda_1+\lambda_2}\right]\tau_1 + \left[\frac{\lambda_2}{\lambda_1+\lambda_2}\right]\tau_2 \quad \textit{and} \quad \lambda = \lambda_1+\lambda_2 \tag{1.14}$$

Note however, that the distribution by (1.13) will not generally have this simple geometric form, as due to the common admissibility restrictions by $\boldsymbol{C}$*.*

1.2.2.3 More examples

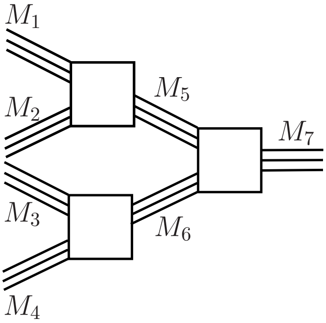


Fig. 1.7: Multiple class circuit switch.

Example 1.2.8 (Circuit Switching) *Example 1.2.3 can directly be extended to multistage switch networks, in which a type-i call requires an available trajectory, that is a free channel from each of its channel groups along its trajectory, such as illustrated in figure 1.7 with the coordinate convex capacity constraints*

$$\begin{cases} m_i \le M_i \quad i = 1,\ldots,4 \\ m_1 + m_2 \le M_5 \\ m_3 + m_4 \le M_6 \\ m_1 + m_2 + m_3 + m_4 \le M_7 \end{cases}$$

Example 1.2.9 (Alternate routing) *Example 1.2.5 can be extended to multiple specialized servers such as naturally arising in communication routes or call center skills provided the 'call packing' principle is assumed, i.e. a job should always uses an available server of its highest preference.*

As a hierarchical routing example in circuit switching, consider a circuit switching communication network as illustrated in figure 1.8 between locations A, B and C.

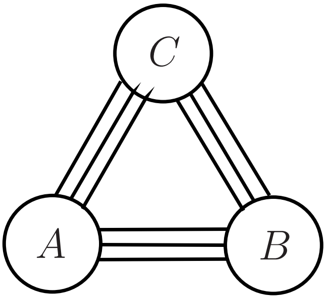


Fig. 1.8: Alternate routing example.

There are input sources for communications between AB, AC and BC with parameters λ_1, λ_2 and λ_3 and with exponential call duration parameters μ_1, μ_2 and μ_3 for each of this call types. The number of links between the locations is limited by N_1 for AB, N_2 for AC and N_3 for BC. A type 2 or type 3 is directly lost if all N_2 or N_3 channels are busy. For type 1 requests between AB, however, alternate routing *is used. That is, if all N_1 links are busy, the AB connection can be made via C, which requires one link between AC and BC at the same time. In addition, call packing is in order. That is, if a direct AB link becomes available, an alternately routed transmission is instantaneously switched to this link.*

With m_r the number of ongoing type r communications, the coordinate convexity conditions now apply with:

$$\begin{cases} m_1 \leq \min(N_1+N_2, N_1+N_3) \\ m_2 \leq N_2 - (m_1 - N_1)^+ \\ m_3 \leq N_3 - (m_1 - N_1)^+ \end{cases}$$

Example 1.2.10 (Two other coordinate convex examples.) *As mentioned in remark 1.2.6, the examples 1.2.3, 1.2.4 and 1.2.5, although different in practical physical descriptions, seem to be identical in its graphical (and mathematical) form.*

Nevertheless, by just the condition of a coordinate convex region $\mathbf{C}$ *(and an implicit assumption of call packing so as to justify a sufficient state description for this region), various other product form examples can so be devised, as illustrated in figures 1.9(a) and 1.9(b).*

For example, figure 1.9(a) could represent hospital departments for type-1 and type-2 patients with B_1 and B_2 beds. In addition, a standby department is available in case of excess. However, this department can only be used for one patient type (e.g. as due to contamination risks), up to a maximum of K_1 type-1 or K_2 type-2 patients.

Also figure 1.9(b) could represent hospital departments for type-1 and type-2 patients. Here it is assumed that each patient requires one nurse up to N_i nurses in department $i = 1,2$. The N_2 nurses of department 2 are assumed to be more 'capable'. Therefore, department 2 even allows up to P_2 patients once all N_2 nurses

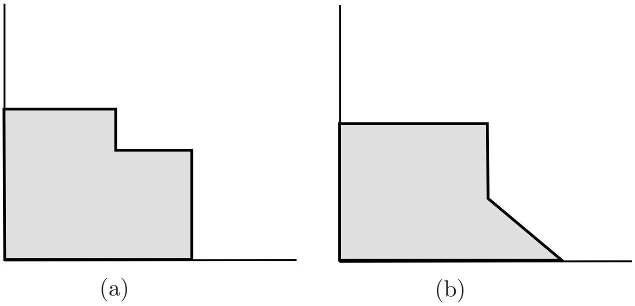


Fig. 1.9: Two other examples with CCP.

are present. In addition, department 1 may also make use of one or more of these specialized N_2 nurses as far as available.

Call packing. As for the 'call packing' assumption in the overflow example 1.2.4, first note that without call packing, that is by naturally assuming that a call remains dedicated to the channel that it is allocated to, one cannot describe the system dynamics by just the numbers m_1 and m_2. In that case, also the separate numbers of channels at the second group have to be kept track of that are occupied by type-1 and type-2 calls. (As the call rates are different and as type 2 calls are lost when all channels at this group are occupied). Say that we would let (p_1, o_1, o_2) denote the state with

$$\begin{cases} p_1 & : \text{type-1 calls at the primary group} \quad \text{(class-1 jobs)} \\ o_1 & : \text{type-1 calls at the secondary group (class-2 jobs)} \\ o_2 & : \text{type-2 calls at the secondary group (class-3 jobs)} \end{cases}$$

Then, without call packing, by regarding the entire system as a multi-class service station, a notion of class balance is necessarily violated for class-2 jobs. For example, for $N_1 = 10$ in state $(4,6,2)$, the outrate due to a class-2 job (type-1 call at group 2) is positive but the inrate into this state due to a class-2 job is equal to 0 (as a type-1 call arrival in state $(4,5,2)$ will lead to state $(5,5,2)$ rather than $(4,6,2)$).

Hence, by relying upon the relationship that a 'product form' solution in numbers of job classes necessarily requires this notion of class-balance, we cannot expect a 'product form' for the overflow example 1.2.4 without the call packing assumption.

With this call packing assumption, in contrast, the product form solution $\pi(m_1, m_2)$ can be concluded as by expression (1.7) with $m_1 = p_1 + o_1$ and $m_2 = o_2$. In fact, this product form also satisfies the class-balances for each job-class $r = 1,2,3$. More precisely, for example 1.2.4 the class balance relations become:

$$\left\{\begin{array}{l}\pi(p_1,o_1,o_2)p_1\mu_1 1_{(p_1>0)}1_{(o_1=0)}+ \\ \pi(p_1,o_1,o_2)(N_1+o_1)\mu_1 1_{(p_1=N_1)}1_{(o_1>0)}+ \\ \pi(p_1,o_1,o_2)o_2\mu_2 1_{(o_2>0)}\end{array}\right\} \begin{array}{l}(1.15.1)\\(1.15.2)\\(1.15.3)\end{array}$$
$$= \qquad (1.15)$$
$$\left\{\begin{array}{l}\pi(p_1-1,o_1,o_2)p_1\lambda_1 1_{(p_1>0)}1_{(o_1=0)}+ \\ \pi(p_1,o_1-1,o_2)\lambda_1 1_{(p_1=N_1)}1_{(o_1>0)}+ \\ \pi(p_1,o_1,o_2-1)\lambda_2 1_{(o_2>0)}\end{array}\right\} \begin{array}{l}(1.15.1)'\\(1.15.2)'\\(1.15.3)'\end{array}$$

Here, to be complete, one must note that the outrate (left hand side) and inrate (right hand side) are indeed both equal to 0 for class-1 jobs when $o_1 > 0$ and hence $p_1 = N_1$. (In that case, the number of class-1 jobs will remain to be N_1 even by a call completion at the primary group). Conversely, an inrate by a class-1 job could not have taken place as it would have come from a state $(N_1, o_1 - 1, o_2)$. Furthermore, also the corresponding outrate for the 'outside station 0' (the total inrate for the system) and inrate (the total outrate from the system) are to be checked.

With $m_1 = p_1 + o_1$ and $\boldsymbol{C}$ as in example 1.2.4, for $(m_1, m_2) \in \boldsymbol{C}$ the class balances $(1.15.i) = (1.15.i)'$ for $i = 1,2,3$, are verified by

$$\pi(p_1,o_1,o_2) = c\frac{1}{m_1!}\left(\frac{\lambda_1}{\mu_1}\right)^{m_1}\frac{1}{m_2!}\left(\frac{\lambda_2}{\mu_2}\right)^{m_2} \qquad (1.16)$$

Example 1.2.11 (Practical use) *As proven and numerically illustrated in [58], this product form expression (1.16) (for the overflow example under call packing as in example 1.2.4) provides secure (and rather accurate) upper bounds for the loss probability of type-1 calls for an overflow system, as in example 1.2.4, without call packing. The call packing principle and its product form consequences, even though 'unrealistic' in particular cases, can thus still be of practical interest, such as for dimensioning purposes to guarantee a sufficiently small loss percentage. (The technical details of the proof in [58] are rather complex and rely upon the Markov reward approach as will be outlined in a chapter later on).*

1.2.3 Job Local Balance: Necessity

1.2.3.1 Introduction: single server system

So far, no distinction or mentioning has been made as to the specific service discipline in order, such as whether jobs have to wait or not and if so whether it is, for example, in first-come first-served or last-come-first-served order. In fact, no other specification is given than just by a simple service rate $\mu(n)$ for all jobs in section 1.2.1 or service rate $\mu_r(m_r)$ for class-r jobs in section 1.2.2.

Apparently, as for the steady state distributions $\pi(n)$ in section 1.2.1 or $\pi(\boldsymbol{m})$ in section 1.2.2, the specific service discipline does not seem to play a role. In contrast, two implicit assumptions have been essential:

- the assumption of just one type of exponential service in section 1.2.1 and
- of an independent service rate for each job-class seperately in section 1.2.2.

In this and the next section we aim to relax both assumptions by allowing different service parameters for different job classes as well as a common service such as by a single server and processor sharing mechanism for different job-classes. As before, let us first obtain some more insight by a simple situation.

Single-server system (FCFS/LCFS). Even though we might only be interested in the total number of jobs present, as the service rates of different jobs are allowed to be different, for either the FCFS or LCFS discipline we necessarily have to keep track of which type of job is in service position 1 and so on. To this end, in a state with n jobs present, let

$$\begin{aligned}&[\boldsymbol{R}] = [r_1, r_2, \dots, r_n] \text{ denote by}\\ &r_i : \text{the job-type of the job at service position } i = 1, 2, \dots, n\end{aligned}$$

In addition, for clarity, in state $[r_1, r_2, \dots, r_n]$ we also use the symbols

$$\begin{aligned}s = r_1 &: \text{for the job-type of the job in service and}\\ l \quad &: \text{for the job-type of the job last entered,}\end{aligned}$$

Hence,

$$l = \begin{cases} r_n \text{ for the FCFS-case}\\ r_1 \text{ for the LCFS-case.}\end{cases}$$

Below we will separately treat the FCFS- and LCFS-served case in order to investigate the existence of a 'product form' or just an 'analytic' solution for $\pi(n)$.

1.2.3.2 Instructive FCFS case

The global balance equations here require that

$$\begin{aligned}&\pi(s, r_2, \dots, r_n)\mu_s + \sum_r \pi(r_1, r_2, \dots, r_n)\lambda_r =\\ &\pi(s, r_2, \dots, r_{n-1})\lambda_l + \sum_r \pi(r, r_1, r_2, \dots, r_n)\mu_r \end{aligned} \tag{1.17}$$

which equate the mathematical out=in rate for any state $[\boldsymbol{R}] = [r_1, \dots, r_n]$. Now note that the physical outrate as due to the job in position 1 has a factor μ_s while the physical inrate is as due to the job at position n has a factor λ_l. As a solution is to

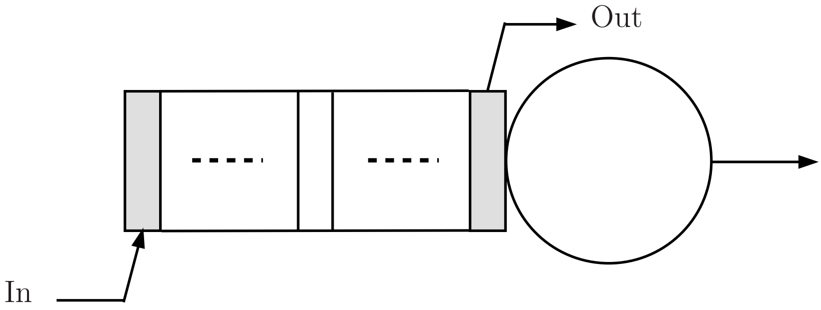


Fig. 1.10: Out and inflow for FCFS-queue.

be sought which holds for any s and l, one cannot expect (find) a simple solution unless the services of different job-types are identical, i.e., for some μ

$$\mu_r = \mu \qquad (r = 1,\ldots,R) \tag{1.18}$$

In other words, without this condition we necessarily seem to fail both a notion of physical out = in rate balance and a notion of out = in rate for each position separately.

Alternatively, under condition (1.18) by substituting $\mu_s = \mu$ and $\mu_r = \mu$ and setting

$$\pi(r_1,\ldots,r_n) = c\prod_{i=1}^{n}\left[\prod_{r}\lambda_r 1_{(r=r_i)}\right]\mu^{-1} \tag{1.19}$$

one easily verifies (1.17) by

$$\begin{aligned} \pi(s,r_2,\ldots,r_n)\mu &= \pi(s,r_2,\ldots,r_{n-1})\lambda_l && \text{(all l)} \\ \pi(r_1,r_2,\ldots,r_n)\lambda_r &= \pi(r,r_1,r_2,\ldots,r_n)\mu && \text{(all r)} \end{aligned} \tag{1.20}$$

The relations (1.20) in turn can again be interpreted as station balance as defined in section 1.2.1. However, even under condition (1.18) a notion of balance for each position (or rather job) separately remains to fail, as necessarily for $n \geq 1$:

the rate out of a state $[r_1,\ldots,r_n]$ due to the job at position $1 > 0$
the rate into this state due to that job at position $1 = 0$

The importance of this notion will become more apparent in the next section. Among other things its failure for the FCFS-situation implies that the solution (1.19) cannot be insensitive, i.e. next to its equality condition (1.18) it also strictly requires exponential services. This will be clarified later on.

1.2.3.3 LCFS-pre case

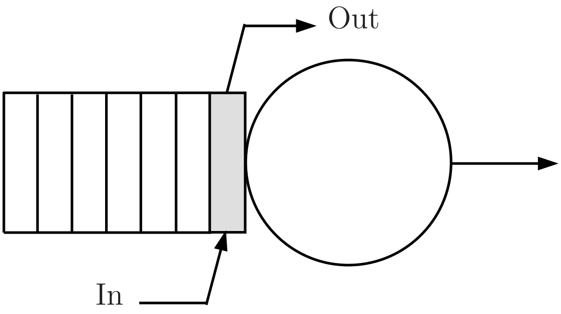


Fig. 1.11: Out and inflow for LCFS-pre queue.

Now consider the Last-come first-served (LCFS) case (with preemption). With the notation from section 1.2.3.2 adopted, in state $[\boldsymbol{R}] = [r_1,\dots,r_n]$ with $s = l = r_1$, the global balance equation here becomes:

$$\pi(s,r_2,\dots,r_n)\mu_s + \sum_r \pi(r_1,r_2,\dots,r_n)\lambda_r =$$
$$\pi(r_2,\dots,r_n)\lambda_s + \sum_r \pi(r,r_1,r_2,\dots,r_n)\mu_r \tag{1.21}$$

Clearly, (1.21) is directly verified by requiring that for any state $\boldsymbol{R} = [r_1,\dots,r_n]$ and with $s = r_1$:

$$\begin{aligned}\pi(s,r_2,\dots,r_n)\mu_s &= \pi(r_2,\dots,r_n)\lambda_s\\ \pi(r,r_1,\dots,r_n)\mu_r &= \pi(r_1,\dots,r_n)\lambda_r\end{aligned} \tag{1.22}$$

with solution:

$$\pi(r_1,\dots,r_n) = c\prod_{i=1}^{n}\prod_{r}\left[\frac{\lambda_r}{\mu_r}1_{(r=r_i)}\right] \tag{1.23}$$

Furthermore, the relations (1.22) have the interpretation that in any state:

> The physical outrate due to a service completion at position 1 =
> the physical inrate due to an arrival at that position 1.

while for any other position $p \neq 1$:

> Both the outrate due to a service completion at that position
> and the inrate due to an arrival at that position are equal to 0.

We have thus verified a notion of balance for each position or job separately. The expression (1.23) in turn appears to factorize to traffic loads for each of these jobs separately. Again, it could therefore be referred to as product form. This property of balance per job (or position) and a detailed ’product form’ in jobs (or positions) separately, thus appear to be interrelated. This will be made more explicit and be extended in sections 1.2.4, 1.2.5 and section 1.3.

First, however, for this particular form of detailed balance, as will be called job-local balance later on, in the next section let us reveal another appealing property that seems to be directly related.

1.2.4 LCFS-pre case: Nonexponential

In this section reconsider the LCFS-preemptive case from section 1.2.3.3 but drop the exponential assumption. From this section onward in the remainder of this chapter, for notational convenience and to distinguish from a subscript for different service stations later on, we will consistently use superscripts to indicate the job-class in order. Hence for job-class r from now on we use:

$$\begin{aligned} \lambda^r &: \text{ as arrival parameter} \\ \mu^r &: \text{ as service parameter in the exponential case} \end{aligned}$$

Instead, assume that class r jobs require an amount of service according to a distribution function:

$$G^r = \sum_{k=1}^{\infty} q^r(k) E(k, \nu^r) \tag{1.24}$$

where $q^r(k)$ represents the probability that the distribution is an Erlang $E(k, \nu^r)$ distribution of k exponential phases with parameter ν^r. Here we refer to remark 1.2.14 below to justify the restriction to this class of distributions. Let

$$\begin{cases} \mu^r = [\tau^r]^{-1} \\ \tau^r = \sum_{k=1}^{\infty} q^r(k)\, [k/\nu^r] \\ H^r(a) = [\tau^r \nu^r]^{-1} \sum_{k=a}^{\infty} q^r(k) \end{cases} \tag{1.25}$$

Hence, μ^r can be seen as equivalent to the exponential parameter for the pure exponential case and τ^r is the mean service requirement. Furthermore, the terms $H^r(\cdot)$, which sum up to 1, can be seen as steady state probabilities for the number of residual exponential phases up to a next renewal in a discrete renewal process with (inter) renewal distribution function G^r. These probabilities satisfy the discrete renewal relation:

$$H^r(a) = H^r(a+1) + H^r(1) q^r(a) \tag{1.26}$$

Now note that the system dynamics or rather the transition rates requires one to keep track of the residual number of exponential phases for each job. Therefore, let

$$[(r_1,a_1),\dots,(r_n,a_n)] \tag{1.27}$$

denote that n jobs are present with the job at position i of class r_i at position i with a_i residual exponential phases of service requirement. The following product form can then be proven.

Result 1.2.12 (Detailed product form)

$$\pi\left([(r_1,a_1),\dots,(r_n,a_n)]\right) = c\prod_{i=1}^{n}\left\{\prod_{r}\left[\frac{\lambda^r}{\mu^r}\right]H^r(a_i)1_{(r=r_i)}\right\} \tag{1.28}$$

Proof. Motivated by section 1.2.3.3 for the exponential case, the proof is based on showing:

$$\begin{aligned}&\text{The rate out of state } [(r_1,a_1),\dots,(r_n,a_n)] \text{ due to the job at position } 1 =\\ &\text{The rate into state } [(r_1,a_1),\dots,(r_n,a_n)] \text{ due to the job at position } 1\end{aligned} \tag{1.29}$$

For clarity, write $r_1 = s$ and $a_1 = a$. In formula, (1.29) then requires

$$\begin{aligned}&\pi\left([(s,a),(r_2,a_2),\dots,(r_n,a_n)]\right)\nu^s\\ &=\\ &\pi\left([(r_2,a_2),\dots,(r_n,a_n)]\right)\lambda^s q^s(a)+\\ &\pi\left([(s,a+1),(r_2,a_2),\dots,(r_n,a_n)]\right)\nu^s\end{aligned} \tag{1.30}$$

By substituting (1.28), this can be rewritten as requiring that

$$\begin{aligned}&\pi\left([(s,a),(r_2,a_2),\dots,(r_n,a_n)]\right)\nu^s =\\ &\pi\left([(s,a),(r_2,a_2),\dots,(r_n,a_n)]\right)\nu^s\left[\lambda^s q^s(a)\frac{1}{\nu^s}\frac{\mu^s}{\lambda^s}\frac{1}{H^s(a)}+\frac{H^s(a+1)}{H^s(a)}\right]\end{aligned} \tag{1.31}$$

With $\mu^s = [\tau^s]^{-1}$, this is satisfied by the renewal relation (1.26).

As the rate out of and into the state $([(r_1,a_1),\dots,(r_n,a_n)])$ due to a departure from and arrival at any other position $p \neq 1$ are both equal to 0, for each position p separately, the notion of job-local balance is verified. In addition, also the total inrate and the total outrate for the system have to be shown to be equal as by:

$$\begin{aligned}&\sum_r \pi\left([(r_1,a_1),\dots,(r_n,a_n)]\right)\lambda^r =\\ &\sum_r \pi\left([(r,1),(r_1,a_1),\dots,(r_n,a_n)]\right)\nu^r\end{aligned} \tag{1.32}$$

and verified (in fact for each job-class r separately), by substituting (1.28)

$$\pi\left([(r,1),(r_1,a_1),\dots,(r_n,a_n)]\right) = \pi\left([(r_1,a_1),\dots,(r_n,a_n)]\right)H(1)\frac{\lambda^r}{\mu^r} \tag{1.33}$$

and using that (also see (1.26)): $H^r(1) = [\tau^r \nu^r]^{-1}$ and $\tau^r = 1/\mu^r$. The global balance equation is hereby satisfied. This completes the proof of the result 1.2.12. □

Result 1.2.13 (Insensitive 'product form' for LCFS-pre case) *For arbitrary service distribution of the form (1.24):*

$$\begin{cases} (1.23)\ \textit{holds and} \\ \pi(n) = c(\lambda\tau)^n\ \textit{with} \\ \tau = \sum_r p^r \tau^r\ ;\ p^r = \lambda^r/\lambda\ ;\ \lambda = \sum_r \lambda^r \end{cases} \tag{1.34}$$

Proof. By using the factorizing form of the product form of result 1.2.12, by summing over all possible numbers of residual phases and recalling that the terms $H^r(\cdot)$ represent renewal probabilities that sum up to 1, first conclude that

$$\begin{aligned} \pi(r_1,\dots,r_n) &= \sum_{a_1,\dots,a_n} \pi([(r_1,a_1),\dots,(r_n,a_n)]) \\ &= c\prod_{i=1}^{n} [\lambda^{r_i}\tau^{r_i}] \left[\sum_{a_i=1}^{\infty} [H^{r_i}(a_i)]\right] \\ &= c\prod_{i=1}^{n} [\lambda^{r_i}\tau^{r_i}] \end{aligned} \tag{1.35}$$

The proof is completed by

$$\pi(n) = c\sum\nolimits_{r_1,\dots,r_n} \pi(r_1,\dots,r_n) = c\left[\sum_r \lambda^r\tau^r\right]^n = c(\lambda\tau)^n$$

□

Remark 1.2.14 (Insensitivity and general nonexponential services) *Result 1.2.13 shows that the steady state distribution $\pi(n)$ is not dependent on the actual service distributions other than by their means τ^r. Such a result is known in the literature as '**insensitivity**'. In fact, as arbitrary nonnegative and continuous distributions can be approximated arbitrarily closely, in weak convergence sense, by mixtures of Erlang distributions, i.e. by distributions of the form (1.24), by weak continuity arguments the closed form (1.34) can also be concluded for arbitrary service distributions.*

Remark 1.2.15 (Insensitivity ⇔ job-local balance) *Note that the insensitivity result 1.2.13 for the total number of jobs has been proven by showing a notion of physical balance for each job (position) separately. This relationship appears to be generally valid, as will also become more apparent in the next section. More precisely, as shown in [47], [48] and [26] in more abstract setting the concepts of insensitivity and, as it will be called here, job-local balance, appear to be one-to-one related.*

Remark 1.2.16 *Chapter 3 by P.G. Taylor provides different insights and results of insensitivity. The interested reader on this intriguing phenomenon is referred to this chapter.*

1.2.5 Symmetric Disciplines and Job-Local-balance (JLB)

In this section we will combine and extend the 'product form' insights and results from sections 1.2.3.2, 1.2.3.3 and section 1.2.4 to a more general setting of service disciplines (directly related to the work by [3], [10], [11], [33], [34]).

To this end, as before, assume that jobs are numbered by service positions $1,\dots,n$ when n jobs are present. The service discipline is characterized by a 3-tuple (f,γ,δ) of functions which represent by:

$$\begin{cases} f(n) & \text{the total service capacity when } n \text{ jobs are present where} \\ & f(n) > 0 \text{ for } n > 0 \\ \gamma(p \mid n) & \text{the fraction of this capacity assigned to the service} \\ & \text{position } p\ ;\ p = 1,\dots,n \text{ when } n \text{ jobs are present} \\ \delta(p \mid n-1) & \text{the probability that an arriving job when } n-1 \text{ jobs} \\ & \text{are present is assigned position } p\ ;\ p = 1,\dots,n > 0. \end{cases}$$

As the service positions are assumed to be successive with only one job in each position, also a shift mechanism is operated. When a job at position p completes its service the jobs at positions $p+1,\dots,n$ are shifted to positions $p,\dots,n-1$. When $n-1$ jobs are present and an arriving job is assigned position p, the jobs previously at positions $p,\dots,n-1$ are shifted to positions $p+1,\dots,n$. As an additional property, that will be distinguished, a service discipline of the form above, is said to be *symmetric* if:

$$\delta(p \mid n-1) = \gamma(p \mid n) \qquad \text{for all } p = 1,\dots,n \text{ and } n > 0 \tag{1.36}$$

The present parametrization covers a reasonably large class of natural service disciplines as:

D1: 1-server FCFS: $f(n) = 1$
$\delta(n \mid n-1) = \gamma(1 \mid n) = 1$ and
$\delta(p \mid n-1) = \gamma(p \mid n) = 0$ otherwise

D2: 1-server LCFS-pre: $f(n) = 1$
$\delta(1 \mid n-1) = \gamma(1 \mid n) = 1$ and
$\delta(p \mid n-1) = \gamma(p \mid n) = 0$ otherwise

D3: 1-server Processor Sharing: $f(n) = 1$ and
$\delta(p \mid n-1) = \gamma(p \mid n) = 1/n$, for all p

D4: Pure multi-server (PS) system: $f(n) = n$ and
$\delta(p \mid n-1) = \gamma(p \mid n) = 1/n$, for all p

Here it is to be noted that the most natural FCFS discipline (D1) is not (and cannot be) parameterized as a symmetric discipline, while the disciplines D2, D3 and D4 do meet the condition (1.36), that is are indeed symmetric.

Consider a given discipline (f, γ, δ) and let jobs arrive and be serviced at the station as in sections 1.2.3 and 1.2.4, that is with different job classes with arrival rate λ^r and exponential service requirements with parameter μ^r for job class r.

As before, let $[\boldsymbol{R}] = [r_1, \ldots, r_n]$ denote the state of job classes for each service position when n jobs are present. Note that we need to keep track of this detailed state description even though we might eventually only be interested in just the total number of jobs. Let

$$F(n) = \left[\prod_{k=1}^{n} f(k)\right]^{-1} \tag{1.37}$$

Result 1.2.17 *Let* $\boldsymbol{S}$ *denote the set of symmetric and* $\boldsymbol{NS}$ *of non-symmetric disciplines. Then for any discipline* $\boldsymbol{D}$ *of the form* (f, γ, δ) *with either one of the two conditions:*

$$\boldsymbol{D} \in \boldsymbol{S} \quad \Leftrightarrow (1.36)$$

$$\boldsymbol{D} \in \boldsymbol{NS} \Leftrightarrow \mu_r = \mu \text{ for all } r$$

we have

$$\pi(\boldsymbol{R}) = c\,F(n) \prod_{p=1}^{n} [\lambda^{r_p}/\mu^{r_p}] \qquad ,\boldsymbol{D} \in \boldsymbol{S} \tag{1.38}$$

$$\pi(\boldsymbol{R}) = c\,F(n) \mu^{-n} \prod_{p=1}^{n} \lambda^{r_p} \qquad ,\boldsymbol{D} \in \boldsymbol{NS} \tag{1.39}$$

Proof. For notational convenience let

$$[\boldsymbol{R} - r_p] = (r_1, \ldots, r_{p-1}, r_{p+1}, \ldots, r_n) \qquad (p = 1, \ldots, n)$$

$$[\boldsymbol{R} + s_p] = (r_1, \ldots, r_{p-1}, s, r_p, r_{p+1}, \ldots, r_n) \; (p = 1, \ldots, n+1)$$

where $r = r_p$ identifies the job-class for the job at position p in state $\boldsymbol{R}$ and s the job-class for the job at position p in state $[\boldsymbol{R} + s_p]$. The global balance equation in state $\boldsymbol{R}$ then becomes

$$\begin{array}{ll} \left\{ \begin{array}{l} \pi([\boldsymbol{R}]) \sum_{p=1}^{n} f(n)\gamma(p \mid n)\mu^r + \\ \pi([\boldsymbol{R}]) \sum_{p=1}^{n+1} \sum_s \lambda_s \delta(p \mid n) \end{array} \right\} & \begin{array}{l} (1.40.1.p) \\ (1.40.2.p) \end{array} \\ = & \qquad\qquad (1.40) \\ \left\{ \begin{array}{l} \sum_{p=1}^{n} \pi([\boldsymbol{R} - r_p])\lambda^r \delta(p \mid n-1) + \\ \sum_{p=1}^{n+1} \sum_s \pi([\boldsymbol{R} + s_p]) f(n+1)\gamma(p \mid n+1)\mu^s \end{array} \right\} & \begin{array}{l} (1.40.1.p)' \\ (1.40.2.p)' \end{array} \end{array}$$

Now distinguish for a symmetric or non-symmetric discipline D. First consider the symmetric case.

***D* : Symmetric**. By assuming the form (1.38) we can write

$$\pi([\boldsymbol{R} - r_p]) = \pi([\boldsymbol{R}])\mu^r f(n)[\lambda^r]^{-1} \tag{1.41}$$

$$\pi([\boldsymbol{R} + s_p]) = \pi([\boldsymbol{R}])\lambda^s[\mu^s f(n+1)]^{-1} \tag{1.42}$$

By substituting (1.41) in $(1.40.1.p)$ and $(1.40.1.p)'$, cancelling equal terms and using (1.36), we directly verify $(1.40.1.p) = (1.40.1.p)'$ separately. Similarly, by substituting (1.42) in $(1.40.2.p)$ and $(1.40.2.p)'$ and using (1.36) we verify $(1.40.2.p) = (1.40.2.p)'$ for each possible class s and position $p = 1, \dots, n+1$.

***D*: Non Symmetric.** We can substitute $\mu^r = \mu$ and $\mu^s = \mu$ in $(1.40.1.p)$ and $(1.40.1.p)'$. By substituting (1.39), both (1.41) and (1.42) remain valid with $\mu^r = \mu$ and $\mu^s = \mu$ substituted. As $\mu^r = \mu$ and $\mu^s = \mu$ can then be taken outside the summations over p in $(1.40.1.p)$ and $(1.40.1.p)'$, (1.40) follows directly by using that

$$\left\{ \begin{array}{l} \sum_{p=1}^{n} \gamma(p \mid n) = \sum_{p=1}^{n} \delta(p \mid n-1) = 1 \\ \sum_{p=1}^{n+1} \delta(p \mid n) = \sum_{p=1}^{n+1} \gamma(p \mid n+1) = 1 \end{array} \right. \tag{1.43}$$

□

Remark 1.2.18 (Insensitivity of symmetric disciplines) *As shown in the proof of result 1.2.17, symmetric disciplines guarantee the notion of job-local balance to be satisfied. In more abstract setting this notion has been shown in [47], [48] and [26] to be both sufficient and necessary for insensitivity. As a consequence, also for symmetric disciplines the product form result can be shown to be insensitive, that is to apply for arbitrary service distributions with mean $\tau^r = 1/\mu^r$ for job-class r. This insensitivity conclusion for symmetric disciplines has first been shown explicitly in [2] and can also be found in [24] and [10]. A straightforward and selfcontained proof in the present setting can be given similarly to the LCFS-pre case as in section 1.2.4. It is omitted here as it is also covered by the generalized setting of invariant disciplines in section 1.3.3.*

By result 1.2.17 and taking remark 1.2.18 into account, as in result 1.2.13 and its proof, we can conclude:

Result 1.2.19 *For arbitrary discipline D of the form (f,γ,δ) with $\mu^r = \mu = \tau^{-1}$ for all r when **D** $\in$ **NS** and for arbitrary service distributions when **D** $\in$ **S**, we have*

$$\begin{cases} \pi(n) = c(\lambda\tau)^n \left[\prod_{k=1}^{n} f(k)\right]^{-1} \quad \textit{with} \\ \tau = \sum_r p^r \tau^r \ ; \ p^r = \lambda^r/\lambda \ ; \ \lambda = \sum_r \lambda^r \end{cases} \tag{1.44}$$

Remark 1.2.20 (Job local balance and product form) *Again, as in the proof for the LCFS-pre case in section 1.2.4, note that the factorizing form of the steady state expression into the traffic ratio's of the individual jobs, relied upon showing balance for each position (job) p, as if these can be regarded as being independent, separately. To distinguish its concept from the class and station balances in section 1.2.2 and section 1.2.1, this most detailed notion of balance will therefore be referred as job-local balance.*

Remark 1.2.21 (Necessity of symmetric discipline) *Conversely, in [25] it has been shown that a notion of job-local balance for disciplines as parameterized in this section, requires the symmetric condition (1.36). In other words, by also referring to the one-to-one relationship between job-local balance and insensitivity, as mentioned and referenced to literature in remark 1.2.15, we can thus conclude that a discipline, which only depends on the total number of jobs n, necessarily has to be symmetric in order to be insensitive.*

1.3 Invariant Disciplines and JLB

1.3.1 Invariance Condition

The disciplines defined in section 1.2.5 only depend on the total number of jobs present and the order of arrival and not on which type of jobs are possessing which positions. By this parametrization, for example, we cannot model different servicing for different job-classes or some kind of priority for one type of job over the other. In this section therefore a more extended discipline is provided that also takes into account which type of jobs have arrived and in which order.

To this end, in state $\boldsymbol{R} = [r_1, r_2, \ldots, r_n]$, where again positions are assumed under the shift protocol as in section 1.2.5, let the functions (f,γ,δ) be represented by:

$f(r_1,r_2,\ldots,r_n)$: is the total capacity that the facility provides

$\gamma(p \mid r_1,r_2,\ldots,r_n)$: is the fraction of this capacity assigned to position p

$\delta(p \mid r_1,r_2,\ldots,r_n)$: is the probability that a job of class $r - r_p$ is accepted and assigned position p when arriving in state $(r_1,\ldots,r_{p-1},r_{p+1},\ldots,r_n)$. Further, we also assume the shift protocol as before.

Remark 1.3.1 (Blocking and service delay) *It is emphasized that we allow*

$$\begin{aligned} &\textstyle\sum_p \delta(p \mid r_1,\ldots,r_n) \leq 1 \\ &\textstyle\sum_p \gamma(p \mid r_1,\ldots,r_n) \leq 1 \end{aligned}$$

In particular, this implies that an arrival can be blocked. *In this case it is assumed to be lost. In addition, a fraction of the service capacity (e.g. by a single server) can also be lost (which can also be regarded as a service thinning or delay).*

Remark 1.3.2 (Definition of δ) *Merely for presentational convenience in the definition of the assignment function δ the arriving job is included. Note that this includes the possible dependence on the job-class of the arriving job.*

Remark 1.3.3 (One-Parallel queues) *(Single queue) Clearly, the parametrization from section 1.2.5 is included by*

$$\begin{cases} f(r_1,\ldots,r_n) = f(n) \\ \gamma(p \mid r_1,\ldots,r_n) = \gamma(p \mid n) \\ \delta(p \mid r_1,\ldots,r_n) = \delta(p \mid n-1) \end{cases} \tag{1.45}$$

(Parallel class-queues) The parametrization also allows to group jobs of the same class in separate registers. More, precisely, in a state with m^r jobs of class r,

$$\begin{cases} \textit{class 1 at positions } 1,\ldots,m^1 \\ \textit{class 2 at positions } m^1+1,\ldots,m^1+m^2 \\ \vdots \\ \textit{class r at positions } m^1+\ldots+m^{r-1}+1,\ldots,m^1+\ldots+m^{r-1}+m^r \end{cases}$$

Position $p^r = m^1+\ldots+m^{r-1}+p$ then corresponds to the p-th position in register r for class r. By setting

$$\begin{cases} f(r_1,\ldots,r_n) = \sum_r f(m^r) \\ \delta(p^r \mid r_1,\ldots,r_n) = \delta^r(p \mid m^r - 1) \\ \gamma(p^r \mid r_1,\ldots,r_n) = \gamma^r(p \mid m^r)[f^r(m_r)/\sum_r f^r(m_r)] \end{cases} \tag{1.46}$$

we can then let each class r have its own discipline $(f^r, \gamma^r, \delta^r)$ as according to section 1.2.5.

Before providing some examples let us directly resent the necessary condition which is required as an extension of the symmetric condition. To this end, we need to introduce some additional notation.

Notation. Let $\boldsymbol{P}$ denote the set of admissible states $\boldsymbol{R} = (r_1, \ldots, r_n)$. As the shift protocol also changes positions of other jobs upon arrival or departure of a job, the actual positions that jobs got upon arrival is not directly readable form a state $(r_1, \ldots, r_n)$. We therefore use the notation of arrival orders $(p_1, \ldots, p_n)$ which indicates that the job at position p_k was the k-th arriving job from the jobs present. From the arrival order $p_1, \ldots, p_{k-1}, p_k$ and the state description $(r_1, \ldots, r_n)$, we then know exactly the job classes $r_{p_1}, r_{p_2}, \ldots, r_{p_k}$ at these positions, that is the job classes in order of their arrival denoted by $(r_{p_1}, r_{p_2}, \ldots, r_{p_k})$. In addition, we also know the position $\bar{p_k}$ that was assigned to the k-arriving job r_{p_k} upon its arrival.

The symmetry condition (1.36) can now be extended to:

Service invariance condition (SIC). For any $\boldsymbol{R} = [r_1, r_2, \ldots, r_n]$ there exists at least one $p \leq n$ such that $\gamma(p \mid r_1, \ldots, r_n) > 0$ and

$$\begin{aligned} &\delta(p \mid r_1, \ldots, r_n) = 0 \Longleftrightarrow \\ &\gamma(p \mid r_1, \ldots, r_n) = 0 \qquad\qquad (p = 1, \ldots, n) \end{aligned} \tag{1.47}$$

Furthermore, there exists a function Ψ such that for any $(r_1, \ldots, r_n) \in \boldsymbol{P}$ and for any permutation $(p_1, \ldots, p_n) \in (1, \ldots, n)$ of arrival orders for which the denominators in the product below are positive:

$$\Psi(r_1, \ldots, r_n) = \prod_{k=1}^{n} \left[\frac{\delta(\bar{p_k} \mid r_{p_1}, \ldots, r_{p_k})}{\gamma(\bar{p_k} \mid r_{p_1}, \ldots, r_{p_k}) f(r_{p_1}, \ldots, r_{p_k})} \right] \tag{1.48}$$

Or equivalently, such that for any $(r_1, \ldots, r_n) \in P$ and $p \leq n$ for which the denominator is positive:

$$\Psi(r_1, \ldots, r_n) = \Psi(r_1, \ldots, r_{p-1}, r_{p+1}, \ldots, r_n) \left[\frac{\delta(p \mid r_1, \ldots, r_n)}{\gamma(p \mid r_1, \ldots, r_n) f(r_1, \ldots, r_n)} \right] \tag{1.49}$$

Remark 1.3.4

1. *(Instantaneous attention) Condition (1.47) reflects a requirement of instantaneous attention.*
2. *(Interpretation of (1.48)) Roughly speaking, the invariance condition (1.48) requires that it should not matter in which order jobs arrive, even though state*

dependence is involved, if we consider $\delta(\cdot \mid \cdot)$ as arrival and $\gamma(\cdot \mid \cdot)f(\cdot)$ as departure rate.

3. *(Reversibility) Condition (1.48) does in fact follows from a so-called principle of reversibility (see Kelly 79). This principle will also come along in extended form of 'adjoint reversibility' in the next section.*
4. *(Condition (1.48) and (1.49)) Either by the principle of reversibility or directly, the equivalence of the conditions (1.48) and (1.49) can so be proven rather easily and is left to the reader.*
5. *(Use of conditions (1.48) or (1.49)) Condition (1.48) can be seen as the condition for its insight whether or not a discipline can be expected to be invariant. Condition (1.49) is the more practical form that will be used in the proof as well as to specify the product form in order.*

1.3.2 Service invariant examples

First let us illustrate that the service invariance condition provides a useful generalization of the standard symmetric case. It includes for example:

- symmetric systems with *class interdependent* blocking
- systems with *class interdependent* servicing
- and combinations

In these examples for a state $(r_1, r_2, \ldots, r_n)$, as before let $\boldsymbol{m} = (m^1, m^2, \ldots, m^R)$ denote by m^r the number of class r jobs present.

Example 1.3.5 (Coordinate convex blocking and symmetric disciplines) *Consider a coordinate convex region $\boldsymbol{C}$, for example in a 2-class case by*

$$\boldsymbol{C} = \left\{(m_1, m_2) \mid m^1 \leq M^1, m^2 \leq M^2, m^1 + m^2 \leq M^1 + M^2\right\}$$

Jobs accepted are serviced by a symmetric discipline (f, γ, δ), e.g. the single or multi server processor sharing discipline D_3 or D_4, satisfying (1.36). The class dependent discipline is parameterized by

$$\begin{cases} \delta(p \mid r_1, \ldots, r_n) = 1_{\boldsymbol{C}}(\boldsymbol{m})\ \delta(p \mid n-1) \\ \gamma(p \mid r_1, \ldots, r_n) = \gamma(p \mid n) \end{cases}$$

The invariance condition is then directly verified with

$$\Psi(r_1, \ldots, r_n) = \begin{cases} F(n) & \textit{as by (1.37) for } (m^1, m^2) \in \boldsymbol{C} \\ 0 & \textit{otherwise} \end{cases}$$

Hence (1.39) applies restricted to $\boldsymbol{C}$.

Example 1.3.6 (Coordinate convex blocking and parallel symmetric disciplines) *The same example also applies with separate symmetric disciplines for each job-class (register) r as by (1.46) in remark 1.3.3 and a coordinate convex common admissibility region* $\boldsymbol{C}$*. In this case, again (1.39) applies but with* $F(n)$ *replaced by:*

$$\Psi(r_1,\ldots,r_n) = 1_{\boldsymbol{C}}(\boldsymbol{m}) \prod_r \left[\prod_{k=1}^{m^r} f^r(k)\right]^{-1}$$

with

$$\delta(p \mid r_1,\ldots,r_n) = (1/m^r)b^r(m^1,m^2) \textit{ for each p with } r_p = r$$
$$\gamma(p \mid r_1,\ldots,r_n) = (1/m^r)\gamma^r(m^1,m^2) \textit{ for each p with } r_p = r$$

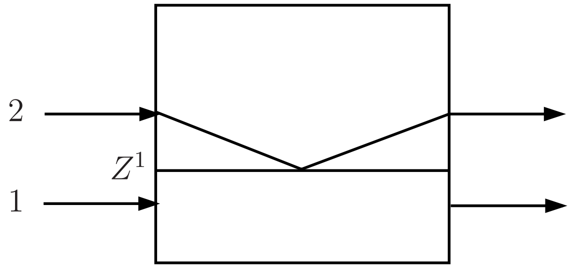


Fig. 1.12: Type-1 dependence for type-2.

Example 1.3.7 (Type 1 level) *Consider a system with 2 job-classes in which class-2 jobs are not accepted or serviced if the number of class-1 jobs is too large, say when* $m^1 \geq Z^1$*, as parameterized by*

$$\delta(p \mid r_1,\ldots,r_n) = \gamma(p \mid r_1,\ldots,r_n) = \begin{cases} 1/n & p = 1,\ldots,n \quad \textit{if } m^1 < Z^1 \\ 1/m^1 & \textit{for any p with } r_p = 1 \textit{ if } m^1 \geq Z^1 \\ 0 & \textit{for any p with } r_p = 2 \textit{ if } m^1 \geq Z^1 \end{cases}$$

Here one may note that for $m^1 \geq Z^1$*, there is no position p at which an arriving job of class 2 is accepted in a state* $(r_1,r_2,\ldots,r_{p-1},r_{p+1},\ldots,r_{n-1})$ *with* $m^1 \geq Z^1$ *jobs of class-1 already present. In other words, a class-2 job is strictly* blocked.

Nevertheless, the state $(r_1,\ldots,r_n)$ *with* $r_p = 2$ *and* $m_1 \geq Z_1$ *is admissible, in which case the servicing of all class-2 jobs is stopped. The invariance condition is directly verified with* $\Psi(\cdot) \equiv 1$

Type 1 jobs can thus be seen as receiving some sort of preference or priority. The servicing sharing, say of a capacity $f(k)$ *when k jobs are present, is processor sharing otherwise. The invariance condition applies with* $\Psi(\cdot) \equiv 1$ *and hence by (1.38)*

$$\pi(\boldsymbol{R}) = c\, F(n) \prod_{r=1}^{2} \left[\frac{\lambda^r}{\mu^r}\right]$$

Example 1.3.8 (Blocking probabilities) *Clearly, by the arrival probability function $\delta(\cdot \mid \cdot)$ blocking probabilities can be incorporated, say by*

$$\delta(p \mid r_1,\ldots,r_n) = b^r(m^r - 1)\delta(p \mid n) \quad \textit{with } r_p = r$$

to represent that a class-r job is accepted only with probability $b^r(k)$ when k jobs of class-r are already present with $b^r(k) > 0$ for $k < M^r$. With $\delta(p \mid n) = \gamma(p \mid n)$ for all $p = 1,\ldots,n$ a symmetric discipline as by (1.36), $f(r_1,\ldots,r_n) = f(n)$ and $F(n)$ given by (1.38), the service invariance condition is satisfied with

$$\Psi(r_1,\ldots,r_n) = F(n) \prod_{r=1}^{R} \prod_{k=1}^{m^r} b_j(k-1) \qquad (m^r \leq M^r \; ; \; r = 1,\ldots,R)$$

Example 1.3.9 (Service scaling) *Alternatively, in line with example 1.3.6 if the number of type-1 jobs becomes too large, say again by $m^1 \geq M^1$, the service capacity might be doubled so as to speed up the servicing while assuming an allover single of multiserver processor sharing servicing as by*

$$\begin{cases} \delta(p \mid r_1,\ldots,r_n) = \gamma(p \mid r_1,\ldots,r_n) = 1/n \;\; p = 1,\ldots,n & \\ f(r_1,\ldots,r_n) = f(n) & \textit{for } m^1 < Z^1 \textit{ while} \\ f(r_1,\ldots,r_n) = 2f(n) & \textit{for } m^1 \geq Z^1. \end{cases}$$

Note that this service acceleration also applies to the type-2 jobs present. In other words, the service speed of one class also depends on the number of the other class present. The service invariance condition is directly verified

$$\Psi(r_1,\ldots,r_n) = F(n)2^{[m^1 - Z^1]^+}$$

Example 1.3.10 (Workload balancing) *As both the arrival and service function δ and γ allow some form of class dependent 'blocking', invariant examples can also be given in which the states can be selectively handled, such as to balance a workload with class preference. For example, again consider a two-class system, say with a single server with arrival and service dependence given by*

$$\delta(p \mid r_1,\ldots,r_n) = \begin{cases} \delta^1(m^1 - 1, m^2)(1/n) \;\; p = 1,\ldots,n \;,\; r_p = 1 \\ \delta^2(m^1, m^2 - 1)(1/n) \;\; p = 1,\ldots,n \;,\; r_p = 2 \end{cases}$$

$$\gamma(p \mid r_1,\ldots,r_n) \; = \gamma^r(m^1,m^2)(1/n) \qquad p = 1,\ldots,n \;,\; r_p = 1,2$$

$$\delta^1(m^1,m^2) = \begin{cases} 1 \\ 1/4 \\ 0 \end{cases} \quad \delta^2(m^1,m^2) = \begin{cases} 0 & m^1 = m^2 - 1 \\ 3/4 & m^1 = m^2 \\ 1 & m^1 = m^2 + 1 \end{cases}$$

$$\gamma^1(m^1,m^2) = \begin{cases} 0 \\ 1/4 \\ 1 \end{cases} \quad \gamma^2(m^1,m^2) = \begin{cases} 1 & m^1 = m^2 - 1 \\ 3/4 & m^1 = m^2 \\ 0 & m^1 = m^2 + 1 \end{cases}$$

Roughly speaking, the system has a triple preference for getting and servicing a class-2 job while in addition the number of jobs of each class are kept to a difference of at most one, as shown in figure

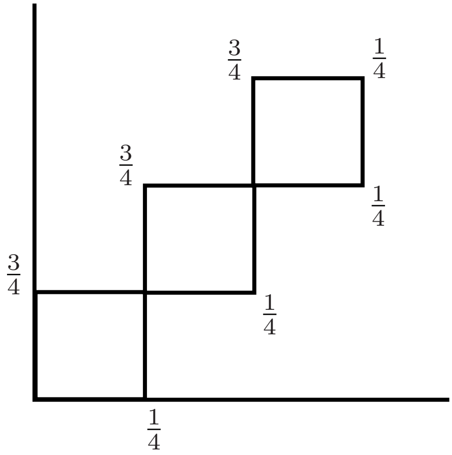


Fig. 1.13: Service Invariance Values Ψ for workload balancing.

The invariance condition is checked with

$$\Psi(r_1,\ldots,r_n) = \begin{cases} 1/4 & m^1 = m^2 + 1 \\ 1 & m^1 = m^2 \\ 3/4 & m^1 = m^2 - 1 \end{cases}$$

Example 1.3.11 (Approximate priority) *Consider a service system which has regular class-1 jbs and which can accommodate (at most) one special class-2 job. Each class-1 job is always serviced at unit rate (as by a multi-server processor sharing discipline). A class-2 job, however, which has a low priority, is only served at a rate τ, and only if no other (class-1) is placed behind it. By letting $\tau \longrightarrow 0$ the discipline thus approximates a strict waiting of the class-2 job and priority for the class-1 jobs. This can be parameterized by*

$$\begin{cases}\delta(n \mid r_1,\ldots,r_n) = 1 & r_n = 2\\ \gamma(n \mid r_1,\ldots,r_n) = \tau/[\tau+(n-1)] & r_n = 2, p < n\\ \delta(p \mid r_1,\ldots,r_n) = 1/[n-1] & r_k = 2, p \neq k\\ \gamma(p \mid r_1,\ldots,r_n) = 1/[n-1] & r_k = 2, k \neq n, p \neq k\\ \gamma(p \mid r_1,\ldots,r_n) = \tau/[\tau+(n-1)] & r_n = 2, p < n\\ \delta(p \mid r_1,\ldots,r_n) = \gamma(p \mid r_1,\ldots,r_n) = 1/n & r_k \neq 2 \, \textit{for all } k \leq n\end{cases}$$

$$\begin{cases} f(r_1,\ldots,r_n) = [\tau+(n-1)] & r_n = 2\\ f(r_1,\ldots,r_n) = [n-1] & r_k = 2 \, \textit{for some } k < n\\ f(r_1,\ldots,r_n) = n & r_k \neq 2 \, \textit{for all } k \leq n\end{cases}$$

The invariance condition now applies with:

$$\Psi(r_1,\ldots,r_n) = \begin{cases} 1/n! & r_k \neq 2 \, \textit{for all } k\\ 1/[\tau(n-1)!] & r_k = 2 \, \textit{for some } k\end{cases}$$

Clearly, an extension to more class-2 jobs is possible along the same type of parametrization.

1.3.3 A generalized symmetric insensitivity result

Now let us extend the results from section 1.2.5 for symmetric disciplines to Service Invariant disciplines. It will be shown that the detailed product form as in sections 1.2.4 and 1.2.5 particularly the related insensitivity results 1.2.13 and 1.2.19 for the simple LCFS-preemptive case and 1.2.19 for symmetric disciplines, can be generalized to class- and position-dependent service disciplines provided it satisfies the invariance condition (1.47). In fact, we will implicitly show that

The service invariance condition $\Longrightarrow$
Job-local balance $\Longrightarrow$
Insensitivity

Formulation. Consider a single service station with arrival rates λ^r and non-exponential service requirements of the form (1.24) for class r-jobs. The service discipline is of the generalized form as in section 1.3.1 under the invariance condition, that is (1.47) and (1.48) or equivalently and (1.49). We further adopt all notation from sections 1.2.4 and 1.2.5. Let

$$[\boldsymbol{R},\boldsymbol{A}] = [(r_1,a_1),(r_2,a_2),\ldots,(r_n,a_n)]$$

be the state which denotes that the job at position p is of class r_p and has a_p residual exponential phases for servicing each with parameter ν_r where $r = r_p$. First, a more technical key-result 1.3.12 is provided. Next, the more practical insensitivity product form result 1.3.13 is concluded.

Result 1.3.12 (Detailed Product From.) *Under the invariance condition (1.47) and with* $\mathbf{R} = (r_1, \ldots, r_n) \in P$*:*

$$\pi([\boldsymbol{R},\boldsymbol{A}]) = c\,\Psi([\mathbf{R}]) \prod_{p=1}^{n} \left\{ \left[\frac{\lambda^r}{\mu^r}\right] \cdot H^r(a) 1_{(r_p=r, a_p=a)} \right\} \tag{1.50}$$

Proof. Again, as in result 1.2.17 and result 1.2.19 for the symmetric discipline, let us first show that a notion of balance is satisfied for each position p separately, i.e.

$$\begin{array}{l}\text{The rate out of this state due to the job at position } p = \\ \text{the rate into this state due to the job at that position.}\end{array} \tag{1.51}$$

Consider a fixed p and the job at position p. For notational simplicity assume $r_p = r, a_p = a$ and introduce the shorthand notation

$$\begin{aligned}&[\boldsymbol{R},\boldsymbol{A}] - (r,a)_p = \\ &((r_1,a_1),\ldots,(r_{p-1},a_{p-1}),(r_{p+1},a_{p+1}),\ldots,(r_n,a_n))\end{aligned}$$

$$\begin{aligned}&[\boldsymbol{R},\boldsymbol{A}] - (r,a)_p + (r,a+1)_p = \\ &((r_1,a_1),\ldots,(r_{p-1},a_{p-1}),(r,a+1),(r_{p+1},a_{p+1}),\ldots,(r_n,a_n))\end{aligned}$$

for the same state with that job left or with its residual service changed from a to $a+1$ phases. Then (1.51) becomes:

$$\begin{aligned}&\pi([\boldsymbol{R},\boldsymbol{A}])\nu^r f(r_1,\ldots,r_n)\gamma(p \mid r_1,\ldots,r_n) = \\ &\pi([\boldsymbol{R},\boldsymbol{A}] - (r,a)_p)\lambda^r \delta(p \mid r_1,\ldots,r_n) q^r(a) + \\ &\pi([\boldsymbol{R},\boldsymbol{A}] - (r,a)_p + (r,a+1)_p)\nu^r f(r_1,\ldots,r_n)\gamma(p \mid r_1,\ldots,r_n)\end{aligned} \tag{1.52}$$

First note that by (1.47):

$$\delta(p \mid r_1,\ldots,r_n) = 0 \Longleftrightarrow \gamma(p \mid r_1,\ldots,r_n) = 0$$

so that (1.52) is trivially satisfied if $\gamma(p \mid r_1,\ldots,r_n) = 0$. Now assume: $\gamma(p \mid r_1,\ldots,r_n) > 0$. By substituting (1.50) we obtain

$$\left\{\begin{aligned}&\frac{\pi([\boldsymbol{R},\boldsymbol{A}] - (r,a)_p + (r,a+1)_p)}{\pi([\boldsymbol{R},\boldsymbol{A}])} = \frac{H^r(a+1)}{H^r(a)} \\ &\frac{\pi([\boldsymbol{R},\boldsymbol{A}] - (r,a)_p)}{\pi([\boldsymbol{R},\boldsymbol{A}])} = \frac{\Psi(r_1,\ldots,r_{p-1},r_{p+1},\ldots,r_n)}{\Psi(r_1,\ldots,r_n)} \left[\frac{\mu^r}{\lambda^r}\right] \frac{1}{H^r(a)}\end{aligned}\right. \tag{1.53}$$

By the invariance condition (1.48) or equivalently (1.49), we can substitute

$$\frac{\Psi(r_1,\ldots,r_{p-1},r_{p+1},\ldots,r_n)}{\Psi(r_1,\ldots,r_n)} = \frac{f(r_1,\ldots,r_n)\gamma(p \mid r_1,\ldots,r_n)}{\delta(p \mid r_1,\ldots,r_n)}$$

where $\delta(p \mid r_1,\ldots,r_n) > 0$ by virtue of (1.47). As a consequence, (1.52) can be reduced to:

$$\begin{cases} \pi([\boldsymbol{R},\boldsymbol{A}])\nu^r f(r_1,\ldots,r_n)\gamma(p \mid r_1,\ldots,r_n) = \\ \pi([\boldsymbol{R},\boldsymbol{A}])\nu^r f(r_1,\ldots,r_n)\gamma(p \mid r_1,\ldots,r_n)\left[\dfrac{\mu^r}{\nu^r}\dfrac{q^r(a)}{H^r(a)} + \dfrac{H^r(a+1)}{H^r(a)}\right] \end{cases} \tag{1.54}$$

With $\mu^r = [\tau^r]^{-1}$ as in the proof of result 1.2.17, the renewal relation (1.26) completes the proof of (1.52), that is, of equality of the outrate and inrate due to the job at any position $p = 1,\ldots,n$. To conclude that global balance is satisfied, it remains to show that:

The outrate due to arrivals=

The inrate due to departures

With $[\boldsymbol{R}+r_p] = (r_1,\ldots,r_{p-1},r,r_p,\ldots,r_n)$, this relation becomes:

$$\begin{aligned} &\pi([\boldsymbol{R},\boldsymbol{A}])\textstyle\sum_r \lambda^r \left[\sum_p \delta(p \mid [\boldsymbol{R}+r_p])\right] = \\ &\textstyle\sum_r\sum_p \pi([\boldsymbol{R},\boldsymbol{A}]+(r,1)_p)\nu^r\gamma(p \mid [\boldsymbol{R}+r_p])f([\boldsymbol{R}+r_p]) \end{aligned} \tag{1.55}$$

By substituting (1.49) and (1.50) again, and noting that $H^r(1) = [\tau^r\nu^r]^{-1} = [\mu^r/\lambda^r]$, we obtain:

$$\frac{\pi([\boldsymbol{R},\boldsymbol{A}]+(r,1)_p)}{\pi([\boldsymbol{R},\boldsymbol{A}])} = \frac{\lambda^r}{\mu^r}\frac{\mu^r}{\nu^r}\frac{\delta(p \mid [\boldsymbol{R}+r_p])}{\gamma(p \mid [\boldsymbol{R}+r_p])f([\boldsymbol{R}+r_p])} \tag{1.56}$$

provided the denominator is positive. By also recalling that $\delta(p \mid [R+r_p]) = 0$ if $\gamma(p \mid [R+r_p]) = 0$ by virtue of (1.47), (1.55) is directly verified by substituting (1.56). This concludes the proof. □

Result 1.3.13 (SIC: Insensitive Product Form) *For arbitrary service distributions with means $[\mu^r]^{-1}$ and for arbitrary service disciplines satisfying the invariance condition, we have*

$$\pi(\boldsymbol{R}) = c\Psi(\boldsymbol{R})\prod_{p=1}^{n}\left[\frac{\lambda^r}{\mu^r}\right]1_{(r_p=p)} \qquad (\boldsymbol{R} \in P)$$

Proof. Identical to that of result 1.2.13 by summing over all possible numbers a_i of residual exponential phases and using that $\sum_a H^r(a) = 1$. □

Remark 1.3.14 (Non service invariant but generalized discipline)
Clearly, as in section 1.2.5 and result 1.2.17, we could also include a detailed product form result for a generalized discipline as in this section without SI condition, provided other stringent conditions as strictly equal and exponential services for all

job-classes are imposed. The situation of independent service mechanisms for each job class separately each with a generalized discipline, also without SI condition, can so be concluded. With general class service interdependence, however, an invariance condition such as also illustrated later on for class balance to hold, over job classes rather than individual jobs as in this section will still be required.

Remark 1.3.15 (Population distribution) *The population distribution $\pi(n)$ is directly obtained by*

$$\pi(n) = \sum_{(r_1,\dots,r_n)\in P} \pi(r_1,\dots,r_n)$$

*In order to get a simple expression for $\pi(n)$, however, the actual form of $\Psi(\cdot)$ and of the set of admissible states **C** play a role.*

1.4 An application, literature discussion and hierarchy review

1.4.1 An $M|G|c|c+m$ application

As shown in section 1.2.3.2 in the single-server case, the natural assumption of a first-come first-served queueing discipline violates a notion of balance for each job or position. More precisely, as by condition (1.47) for the service invariance condition, the notion of job-local balance necessarily requires the condition of instantaneous attention: if accepted a job should also immediately receive an amount of service by which it may complete its service. As a consequence:

> *Any system in which an arriving job may have to wait necessarily fails to satisfy job-local balance and thus as in section 1.3.3 cannot (be expected to) have a simple and 'insensitive product form type expression'.*

This holds for the most simple $M|G|c|c+m$-system, thus with c servers and m waiting places, for any $m > 0$, as in section 1.2.5 note that no symmetric discipline can be defined to cover a FCFS waiting discipline.

Modification. Intuitively, however, this failure of job-local balance for an $M|G|c|c+m$-system can be repaired by simply not allowing waiting positions. To this end, we can either

add extra servers for each waiting position or
delete the waiting positions

that is by modifying the system into an

- $M|G|c+m|c+m$ or
- $M|G|c|c$-system

These pure multi-server loss systems can be parameterized by a symmetric (and thus also invariant) processor sharing discipline (as in section 1.2.5), as well-known and proven in section 1.3.3, these systems have an 'insensitive product form' solution as in result 1.2.19 (or result 1.3.13). Particularly, the corresponding loss probabilities are well-known to be insensitive as Erlang's loss expression for s servers as:

$$\boldsymbol{F}(s) = [(\lambda\tau)^s/s!]\,/\left[\sum_{k=0}^{s}(\lambda\tau)^k/k!\right]$$

Simple bounds. Intuitively, by adding servers we will increase the system capacity and thus decrease the loss probability, while conversely by deleting waiting positions we decrease the system capacity and thus increase the loss probability. We have thus obtained a lower bound $\boldsymbol{B}_L$ and upper bound $\boldsymbol{B}_U$ on the loss probability $\boldsymbol{B}$ by:

$$\begin{aligned}\boldsymbol{B}_U &= \boldsymbol{F}(s)\\ \boldsymbol{B}_L &= \boldsymbol{F}(s+m)\end{aligned}$$

In combination with the inequality:

$$\frac{(\rho-1)^+}{\rho} \leq \boldsymbol{B} \leq \frac{\rho}{(\rho+1)} \tag{1.57}$$

where $\rho = \lambda\tau/s$ as proven by Heyman (1980) and Sobel (1980) (see [60] for the references), and the observation that $F(s) \leq \rho/(\rho+1)$ for any s, the following simple bounds are thus concluded:

$$\max\left[\frac{(\rho-1)}{\rho},\ \boldsymbol{F}(s+m)\right] \leq \boldsymbol{B} \leq \boldsymbol{F}(s) \tag{1.58}$$

Practical relevance and numerical results. Clearly, as adding or deleting servers is a drastic system modification, one cannot expect accurate bounds. However, as the bounds are insensitive and most easily computed, they can be useful as quick secure estimates for the order of magnitude as well as for qualitative purposes as in the optimal design application below (the numerical results below support these claims which show a significant improvement over (1.57) for small traffic values and more accurate intervals for large traffic situations. Here we used $\mu = 1$ and the value $\boldsymbol{B}$ applies to the exponential case.

An optimal design example. Let us give a simple illustration of how these insensitive bounds can be used for qualitative purposes in practical situations. The numbers in this example are chosen rather arbitrarily.

Consider a service station which accommodate at most 10 jobs in total. The total number of servers s, however, is yet unspecified and is to be fixed. Each server incurs a salary cost of 100 dollar per hour. Each lost arrival, in turn, is seen as an opportunity loss of 100 dollar. On hourly basis $\lambda = 10$ and $\mu = 2$. The figure below graphically illustrates the total costs depending on s, the number of servers, as corresponding to (1.58).

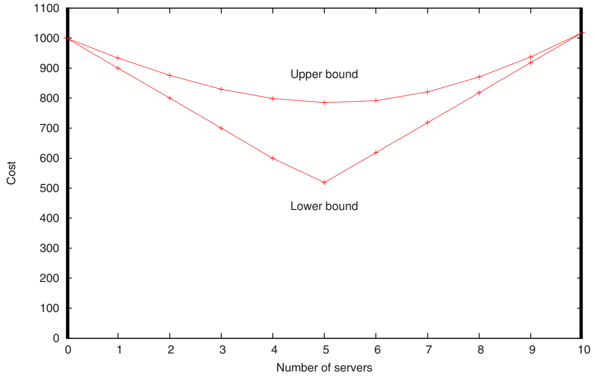


Fig. 1.14: $M|G|c|c+m$ - cost bounds by (1.58).

Since both the lower and upper bound calculations lead to the same optimal number of servers $s = 5$, this number is also most likely to be optimal for the original system, regardless of the distributional forms of the service requirements. Although a 100% guarantee cannot be given, it is the best indication that one can get without further knowledge of the service distributions and approximate calculations. Moreover, as graphically illustrated, in any case the optimal number will be restricted to the '*optimal region*': (3,4,5,6,7), regardless of service distributional form.

1.4.2 Literature discussion

Closed form expressions for queues and queueing networks are generally known to be related to notions of 'partial balance'; that is, by which the global (Kolmogorov) equations are satisfied in some special decomposed form. Most notably, ever since the pioneering work by Erlang in the twenties (e.g. see [9], [43]), birth-death equations for $M|M|s|s+m$-queues are standardly used as starting point in virtually any introductory OR-textbook.

Nevertheless, the *necessity* of these equations rather than just *sufficiency*, as well as its physical rather than mathematical interpretation, as presented in section 1.2.1 as '*station balance*', seems far less commonly emphasized. For its explicit form (1.4) as a product form result for the machine-repair system, as historically known as an Engset system (see [[9], [43]]), a similar statement applies.

Coordinate convex multiple class extensions in the setting of a single service stage with blocking as in section 1.2.2 date back to the seventies as by [15], [31], [39].

Coordinate convex examples as examples 1.2.3 and 1.2.8 can be found in these references. Also the call packing principle and its closed form expressions as in examples 1.2.4 and 1.2.9 are long known in teletraffic literature (e.g. [20], [27], [43]).

The example 1.2.5, its observations in remark 1.2.6 as well as the example 1.2.10 seem to have remained úneported. The special call packing application in example 1.2.11 in order to provide simple bounds for únsolvable overflow systems is based on [58].

Also in the setting of networks with multiple service stages but essentially without blocking (or accessibility constraints), closed (product) form results for multiple job-class extensions have been reported more or less at same time ([3], [10], [11], [32], [33], [44] as will also be referred to in section 1.8.2 of **B**. These references also introduced the condition of symmetric disciplines and proved its relationship with (sufficiency for) product forms, as presented in section 1.2.5.

(The necessity of a discipline to be symmetric (or invariant) to guarantee a notion of balance for each position separately (job-local balance), and correspondingly in order for a discipline to be insensitive, has been shown in [25]).

An extension of these symmetric disciplines related to but more restricted than the service invariance disciplines in section 1.3.1, can already be found in [3], [10], [11]. The conditions in these references are more restricted in that it excludes access or service blocking as covered herein (see remark 1.3.1). The service invariance condition and its insensitive product form relationship as presented in sections 1.3.1 and 1.3.3, but again without blocking, are essentially based upon [24]. The examples 1.3.8-1.3.11 rely upon [25].

The optimal design application in section 1.4.1 and a formal proof of the bounds (1.57) for $M|G|s|s+m$-systems, as by sample path comparison, have been given in [60].

As mentioned, different notions of 'partial' balance and its relationship with a (possibly insensitive) closed (product) form expression have been reported in the literature, as local balance ([3], [10], [11], [47], [48], detailed balance ([32], [33], [34] and job-local balance ([25], [26]). In particular these notions were used in these references in relation to insensitivity (also see [2], [13], [20], [24], [26], [47], [48], [65]). For a more detailed exposition and related references on the phenomenon of insensitivity the reader is also referred to the chapter by *Taylor*.

As such the three notions as used in (this first part of) this chapter are not exclusive or absolute, but simply used for their distinction in line with their natural interpretation and corresponding hierarchy, as will be reviewed in the next section.

1.4.3 A hierarchy review

On the basis of just a single service station rather than on that of a network of service stations, this first part for just a single station (**A**) aimed to illustrate and highlight a number of aspects, which are just as representative for networks of queues, related to the concept of product forms. These aspects are:

1. The notion of a 'product form' as factorizing to different components (e.g. separate stations, different job-classes or individual jobs).
2. Its direct relationship with a form of partial balance with the interpretation of a physical out = in rate for that particular component.
3. The different detailed levels of a product form as determined by its state description.
4. The corresponding system conditions (e.g. on a service discipline or blocking) and service assumptions (as indistinguishable and exponential or not) that might be required.
5. A hierarchy of product form results (as summarized in table 1.1) from:
    - A simple expression, say for just the total numbers of jobs, with hardly no discipline limitation on the one hand but a strict assumption of indistinguishable and exponential services on the other, up to:
    - A most detailed expression which allows distinguishable jobs and which might even apply to arbitrary services on the one hand but only under more restricted system mechanisms (as a symmetric discipline or even rather specialized as an invariant discipline).

Table 1.1: Balance Hierarchy Scheme.

State	**State**	**System**	**Service**	**Product Form**
Global Detailed	Station Class Job local	Stronger Conditions	Stronger Conditions	Stronger Insensitive-PF Result

This scheme is not complete (for example other balance notions not mentioned but which could have fitted in are:

- Cluster balance (at a level of multiple stations, also see section 1.7)
- Group balance (at a level for groups of jobs that move simultaneously)(e.g. [7])

- Source balance (related to job-local balance for the situation of networks in which each job is generated by a specific source, e.g. as in the machine-repair system of section 1.2.1).

Nevertheless, the scheme is meant to be illustrative for the hierarchical (weaker and stronger) results and conditions related to product form expressions as will even be more complicated but with the same flavour in the setting of networks of service stations. Numerous specific product form results that fit within such a scheme have been reported widely in the literature and under different terminologies (as local, detailed or partial balance with a specific meaning). As such the balance notions as used in this chapter are not assumed or claimed to be as unique definition. They are simply used as 'natural terms' for the distinctions in physical interpretations as used in this chapter. This interpretation can be useful, as will be illustrated in the next section, to recognize whether a product form can be expected or not and of what form.

To summarize, the question whether a system has a product form or not might not be easily answered. It may require a more specified formulation such as at which level and under what conditions. And even so the answer might not be as simplistic as it seems.

Particularly, as mentioned, the specific notion of partial balance might be highly practical to obtain more insight in its answer. This insight might lead to either of three practical directions:

- To conclude a product form of specific form
- To conclude that the notion of partial balance necessarily fails so that a corresponding product form cannot exist
- To suggest appropriate product form modifications that might still be practical (for approximate or bounding purposes)

In a second part (**B**), this essential role of a specific form of partial balance in this case of just station balance, and its 'practical' consequences will be illustrated and investigated further for situations with, consecutive service stations and practical features as blocking or service sharing.

B: Product Forms: Tandem and Cluster Structures

1.5 Tandem Queues

1.5.1 Introduction

So far the characteristic feature of a queueing network, rather than just a single service station, has not yet been covered explicitly, that is:

> *Two (or more) successive service stages for a job to be processed, at different service stations. At each of these stations this job may interact with a different set of other jobs.*

Example 1.5.1 (Machine-repair example revisited) *In fact, in line with and as a slight modification of the Engset or Machine-repair system from example 1.2.2 from section 1.2.1, let us first consider a most simple example with blocking.*

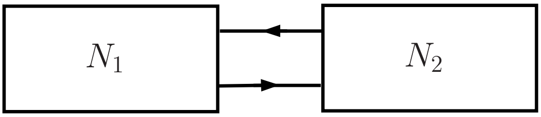


Fig. 1.15: Finite Machine-repair System.

This concerns a closed system with M jobs and two single server stations, say each with a single server with exponential service parameter μ_i, and a routing from one station to the other back and forth. In addition, however, each of these stations has a finite capacity to accommodate at most N_i jobs, $i = 1,2$. When station i is saturated ($n_i = N_i$) jobs from the other station are blocked so that effectively the service of the other can be seen as being 'stopped' as long as the other remains saturated.

Although it is sufficient (as M is fixed) to only specify the number of jobs at one station, let $\boldsymbol{n} = (n_1,n_2)$ denote the number of jobs n_i at either station $i = 1,2$. The global balance equations then become

$$\begin{aligned}
&\left\{\begin{array}{l}\pi(n_1,n_2)\mu_1 1_{(n_1>0)} 1_{(n_2<N_2)} + \\ \pi(n_1,n_2)\mu_2 1_{(n_2>0)} 1_{(n_1<N_1)}\end{array}\right\} && \begin{array}{l}(1.59.1)\\(1.59.2)\end{array} \\
&= && \qquad (1.59)\\
&\left\{\begin{array}{l}\pi(n_1-1,n_2+1)\mu_2 1_{(n_1>0)} 1_{(n_2<N_2)} + \\ \pi(n_1+1,n_2-1)\mu_1 1_{(n_2>0)} 1_{(n_1<N_1)}\end{array}\right\} && \begin{array}{l}(1.59.1)'\\(1.59.2)'\end{array}
\end{aligned}$$

Clearly, as the indicator values that takes into account the finite capacity N_i in the left and right hand side of (1.59.1) for station $i = 1$ and (1.59.2) for station

$i = 2$ are identical, one directly verifies the station balances $(1.59.1)=(1.59.1)'$ and $(1.59.2)=(1.59.2)'$, and thus the global balance (1.59), by the solution.

$$\pi(n_1,n_2) = c\left[\frac{1}{\mu_1}\right]^{n_1}\left[\frac{1}{\mu_2}\right]^{n_2} \qquad (n_1 \le N_1 \; ; \; n_2 \le N_2) \tag{1.60}$$

This example seems to suggest that the notion of station balance also allows capacity constraints and interactions of service stations and ensures a product form solution also in the situation of multiple stations. However, this particular example can in fact still be analyzed by only keeping track of the number of jobs, say $n = n_2$, at one station, that is as a one dimensional system, as if it can be regarded (as in section 1.2.1) as a single (birth-death) service station.

As a first and most simple situation which strictly requires a multi-dimensional description, in the next example therefore, a two station tandem queue is considered (which can also be regarded as equivalent to a closed three station network with station 0 representing the outside).

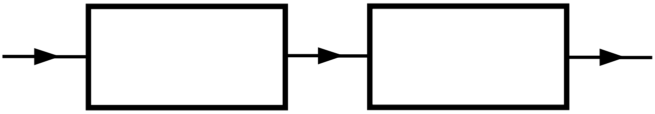


Fig. 1.16: Tandem Queue.

Example 1.5.2 (A tandem queue) *Consider an open system of two station tandem queue with arrival rate λ, and two single server stations in series with exponential service rates μ_i at station $i = 1,2$ with $\lambda/\mu_i < 1$, $i = 1,2$. With $\boldsymbol{n} = (n_1,n_2)$ denoting the number of jobs n_i at station $i = 1,2$, the global balance equations become:*

$$\left\{\begin{array}{l} \pi(n_1,n_2)\lambda + \\ \pi(n_1,n_2)\mu_1 1_{(n_1>0)} + \\ \pi(n_1,n_2)\mu_2 1_{(n_2>0)} \end{array}\right\} \begin{array}{l} (1.61.0) \\ (1.61.1) \\ (1.61.2) \end{array}$$
$$= \tag{1.61}$$
$$\left\{\begin{array}{l} \pi(n_1,n_2+1)\mu_2 + \\ \pi(n_1-1,n_2)\lambda 1_{(n_1>0)} + \\ \pi(n_1+1,n_2-1)\mu_1 1_{(n_2>0)} \end{array}\right\} \begin{array}{l} (1.61.0)' \\ (1.61.1)' \\ (1.61.2)' \end{array}$$

Clearly, this equation is directly verified by each of the more detailed balances $(1.61.i)=(1.61.i)'$ separately by substituting

$$\pi(n_1,n_2) = c\left[\frac{\lambda}{\mu_1}\right]^{n_1}\left[\frac{\lambda}{\mu_2}\right]^{n_2} \qquad \left(\frac{\lambda}{\mu_i} < 1 \; ; \; i = 1,2\right) \tag{1.62}$$

The relations $(1.61.i)$=$(1.61.i)'$ *in turn can be seen as* **station balance** *relation, as formulated as by equating the physical outrate = physical inrate, for*

station 0 as representing the outside for $i = 0$
station 1 for $i = 1$, *and*
station 2 for $i = 2$

The notion of station balance thus seems to be directly responsible for a factorization to the stations as if these are completely independent. In fact, as the state space is unlimited also the normalizing constant c factorizes as $c = (1-\rho_1)(1-\rho_2)$ *so that* $\pi(n_1,n_2) = \pi_1(n_1)\pi_2(n_2)$ *with* $\pi_i(n_i)$ *the steady state distribution of a single server queue.*

However, in this case no interaction between jobs at all is involved, say as due to a finite capacity constraint. To also include these constraints, in the next example let us first assume just a finite capacity constraint at station 2.

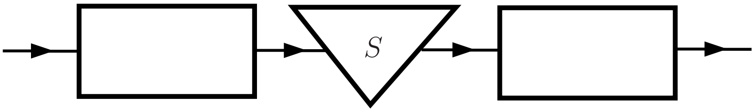


Fig. 1.17: Tandem Queue with Finite Buffer.

Example 1.5.3 (A simple finite tandem queue) *Reconsider the tandem system from example 1.5.2 at which the number of jobs at station 2, such as by a finite (intermediate) storage buffer of size S, is restricted (the job in service included) to a number* $n_2 \leq N_2 = S+1$. *In order for station balance to apply, now note that relation* $(1.61.2) = (1.61.2)'$ *for station 1 would have to be replaced by*

$$\pi(n_1,n_2)\mu_1 1_{(n_1>0)} 1_{(n_2<N_2)} = \pi(n_1-1,n_2)\lambda 1_{(n_1>0)} \qquad (1.63)$$

But clearly, for $n_2 = N_2$, *this relation cannot be satisfied as the left hand side is equal to 0 while the right hand is positive. Similarly,* $(1.61.1) = (1.61.1)'$ *can no longer be satisfied. In other words, for the more natural situation with a finite capacity constraint for the second station, station balance is necessarily violated so that a product form can no longer be expected.*

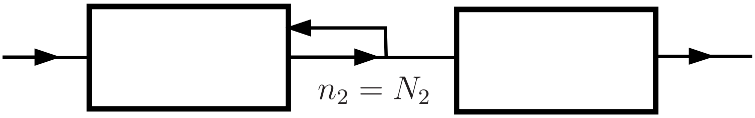


Fig. 1.18: Finite Tandem Queue.

Product Form Modification. This observation, however, can still be practically useful. By artificially assuming that arrivals are blocked when the second station is saturated ($n_2 = N_2$), intuitively both the outrate and inrate for station 1 have become 0 so that (1.61) seems to be restored.

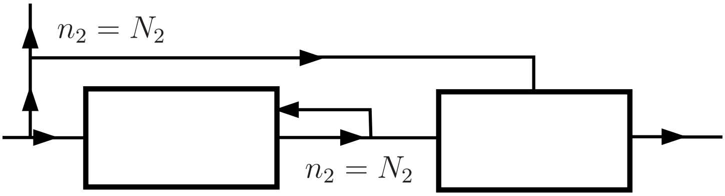


Fig. 1.19: Product Form Modification.

Indeed, under this modification the global balance equations become:

$$\begin{array}{ll}\left\{\begin{array}{l}\pi(n_1,n_2)\mu_1 1_{(n_1>0)}1_{(n_2<N_2)}+\\ \pi(n_1,n_2)\mu_2 1_{(n_2>0)}+\\ \pi(n_1,n_2)\lambda 1_{(n_2<N_2)}\end{array}\right\} & \begin{array}{l}(1.64.1)\\(1.64.2)\\(1.64.3)\end{array}\\ = & \\ \left\{\begin{array}{l}\pi(n_1-1,n_2)\lambda 1_{(n_1>0)}1_{(n_2<N_2)}+\\ \pi(n_1+1,n_2-1)\mu_1 1_{(n_2>0)}+\\ \pi(n_1,n_2+1)\mu_2 1_{(n_2<N_2)}\end{array}\right\} & \begin{array}{l}(1.64.1)'\\(1.64.2)'\\(1.64.3)'\end{array}\end{array} \qquad (1.64)$$

These are directly verified again by the station balance equations $(1.64.i) = (1.64.i)'$ for $i = 1,2,3$ separately by substituting the product form (1.62) restricted to $\{(n_1,n_2) \mid 0 \leq n_1\ ;\ 0 \leq n_2 \leq N_2\}$.

Such a modified product form result, in turn, as based on its required partial balance, can still be useful such as to provide an approximate order or a secure bound for some performance measure of interest. This will be elaborated upon more formally in a separate chapter on bounds and error bounds later on.

In the next section, therefore, the possible existence of product forms, even under 'un'natural system protocols will be explored further for multiple service stations with dependencies such as due to finite capacity restrictions (blocking) or common service sharing. In section 1.7 it will also be explored and numerically illustrated for networks with groups of stations (clusters) having finite constraints.

1.5.2 Product Form Tandem Queues

Consider an open tandem structure of two service stations $i = 1,2$. We aim to investigate the existence of a product form which may possibly include

- Acces blocking
- Stations to be fully congested
- A load dependent service sharing over the stations

To this end, we will allow (but also may need) a state dependent parameters for servicing and routing. To this end, as before let $\boldsymbol{n} = (n_1, n_2)$ denote the number of jobs at stations 1 and 2 and let $\boldsymbol{n} + e_i$ denote the same state except for one job more at station i and $\boldsymbol{n} - e_j$ for one job less at station j. Furthermore, for unification the index $i = 0$ or $j = 0$ is used to represent the 'outside from the system' as a 'station 0' and the convention is used that $\boldsymbol{n} + e_0 = \boldsymbol{n} - e_0 = \boldsymbol{n}$. Jobs arrive at station 1 by a Poisson arrival rate λ and require an exponential amount of service at station 1 and 2 with parameter μ_1 and μ_2. Then, in state $\boldsymbol{n}$, the system dynamics is parameterized by two functions $f_i(\boldsymbol{n})$ and $b_j(\boldsymbol{n})$ as representing:

$f_i(\boldsymbol{n})$: the total service capacity of station $i, i = 1,2$. This capacity can be equal to 0 which represents that that station is effectively stopped

$b_j(\boldsymbol{m})$: the probability that an entering request at station j, that is a transition from station $i = j - 1$ into station j with underlying state $\boldsymbol{m}$, hence from $\boldsymbol{m} + e_i$ into $\boldsymbol{m} + e_j$ is accepted, with

$$\begin{cases} i = j - 1 = 0 & \text{for an arrival at the system and} \\ j = i + 1 = 0 & \text{for a departure from the system} \end{cases}$$

Remark 1.5.4 (Separate service and blocking function) *Clearly, the functions $f_i(\boldsymbol{n})$ and $b_j(\boldsymbol{n})$ can be combined into a single mathematical function*

$$\mu_{ij}(\boldsymbol{n}) = f_i(\boldsymbol{n}) b_j(\boldsymbol{n} - e_i) \qquad (j = i + 1)$$

Nevertheless, a distinction in a separate service and blocking function is made

- *for clarity of its physical interpretations related to possible applications*
- *for the insight in the possible existence of a product form more and*
- *to highlight another characterization of a product form.*

Remark 1.5.5 (Blocking) *With probability $[1 - b_j(\boldsymbol{m})]$ a transition from state $\boldsymbol{n} + e_i$ into $\boldsymbol{n} + e_j$ with $j = i + 1$ will thus be blocked. For an arrival $(i = 0)$ this effectively means that the arrival is lost. For a service completion $(i = 1,2)$ this effectively means that the state remains unchanged (that is $\boldsymbol{n} + e_i$) as if the blocked job will have to undergo a new service at station i. Alternatively, this probability can also be regarded as if the effective service at station i is completely stopped (when this factor is 0 as by complete blocking) or delayed by this factor (when it is positive).*

Remark 1.5.6 (Strict 0-values) *In contrast with results in the literature for processor sharing systems, (e.g. see [4], [10], in which capacity functions are assumed to be strictly positive), it is noted again that the capacity functions can take on 0-values in which the service at a station is stopped. This relaxation is included as it may either arise naturally so as to give full service priority to one station or as it may be required, in order to conclude a product form, even though unnatural as in example 1.5.3. For similar reasons also arrivals may (have to) be blocked. Examples for both situations (natural and unnatural) will be given in sections 1.5.3-1.5.5.*

Station balance and adjoint reversibility. Let $\boldsymbol{C}$ be the set of admissible states. Under the assumption (of ergodicity) for its existence let $\pi(\boldsymbol{n})$ denote the steady state distribution at $\boldsymbol{C}$ as determined by the global balance equations. These require that for any $n \in \boldsymbol{C}$:

$$\begin{array}{ll}\left\{\begin{array}{l}\pi(\boldsymbol{n})\lambda b_1(\boldsymbol{n})+ \\ \pi(\boldsymbol{n})\mu_1 f_1(\boldsymbol{n})1_{(n_1>0)}b_2(\boldsymbol{n}-e_1)+ \\ \pi(\boldsymbol{n})\mu_2 f_2(\boldsymbol{n})1_{(n_2>0)}b_0(\boldsymbol{n}-e_2)\end{array}\right\} & \begin{array}{l}(1.65.0)\\(1.65.1)\\(1.65.2)\end{array} \\ = & \qquad (1.65) \\ \left\{\begin{array}{l}\pi(\boldsymbol{n}+e_2)1_{(\boldsymbol{n}+e_2\in \boldsymbol{C})}\mu_2 f_2(\boldsymbol{n}+e_2)b_0(\boldsymbol{n})+ \\ \pi(\boldsymbol{n}-e_1)1_{(n_1>0)}1_{(\boldsymbol{n}-e_1\in \boldsymbol{C})}\lambda b_1(\boldsymbol{n}-e_1)+ \\ \pi(\boldsymbol{n}-e_2+e_1)1_{(n_2>0)}1_{(\boldsymbol{n}-e_2+e_1\in \boldsymbol{C})}\mu_1 f_1(\boldsymbol{n}-e_2+e_1)b_2(\boldsymbol{n})\end{array}\right\} & \begin{array}{l}(1.65.0)'\\(1.65.1)'\\(1.65.2)'\end{array}\end{array}$$

(Here it is noted that some of the notation and implicit assumptions can be overlapping. For example, if $\boldsymbol{m}+e_2 \notin \boldsymbol{C}$ necessarily $\pi(\boldsymbol{m}+e_2) = 0$ as the state $\boldsymbol{m}+e_2$ is not admissible. Nevertheless, the various functions are used to keep the 'boundary aspects' explicit).

One cannot expect an analytic solution for (1.65) unless for each $i = 0,1,2$ separately we can verify the station balance equation: $(1.65.i) = (1.65.i)'$; that is by a balance of the departure and arrival rate each station and in the natural flow direction of the system dynamics.

With $\mu_0 = \lambda$ and $f_0(n) \equiv 1$, and $i-1 = 0$ for $i = 1$ and $i+1 = 0$ for $i = 2$, the equations $(1.65.i)$ in turn can be rewritten as requiring that for any underlying configuration $\boldsymbol{m}$ (not necessarily in $\boldsymbol{C}$) and $i = 0,1,2$:

$$\left\{\begin{array}{l}\pi(\boldsymbol{m}+e_i)\mu_i f_i(\boldsymbol{m}+e_i)b_{i+1}(\boldsymbol{m})1_{(\boldsymbol{m}+e_i\in C)} = \\ \pi(\boldsymbol{m}+e_{i-1})\mu_{i-1}f_{i-1}(\boldsymbol{m}+e_{i-1})b_i(\boldsymbol{m})1_{(\boldsymbol{m}+e_{i-1}\in C)}\end{array}\right. \qquad (1.66)$$

Here (1.66) is equivalent to the station balances:

$$\begin{array}{lll}(1.65.0) = (1.65.0)' & \text{for } \boldsymbol{m} = \boldsymbol{n} & \text{and } i = 0 \\ (1.65.1) = (1.65.1)' & \text{for } \boldsymbol{m} = \boldsymbol{n}-e_1 & \text{and } i = 1 \\ (1.65.2) = (1.65.2)' & \text{for } \boldsymbol{m} = \boldsymbol{n}-e_2 & \text{and } i = 2\end{array}$$

Before continuing, note that (1.66) already imposes implicit conditions as $\boldsymbol{C}$ is not assumed to be of any particular form (such as coordinate convex as in section 1.2.2).

A non-coordinate convex region $\boldsymbol{C}$ might thus be allowed, as will be illustrated later on (e.g. see example 1.5.18 and 1.5.20). In contrast, though, (1.66) does require compensation by 0-values for the service or blocking functions f and b. For example, if

$$\boldsymbol{m}+e_1 \in \boldsymbol{C} \text{ but also } f_1(\boldsymbol{m}+e_1)b_2(\boldsymbol{m}) = 0,$$

so that the outrate in state $\boldsymbol{m}+e_1$ due to a departure at station 1 = 0, also the inrate into this state $\boldsymbol{m}+e_1$ due to an arrival at station 1 should be equal to 0, by either

$$\boldsymbol{m} \notin \boldsymbol{C} \quad \text{or} \quad f_0(\boldsymbol{m})b_1(\boldsymbol{m}) = b_1(\boldsymbol{m}) = 0.$$

Now in order to investigate the existence of a solution for this more restricted (station) balance relation (1.66), define a continuous-time Markov chain, which will be called the adjoint Markov chain, at the same state space $\boldsymbol{C}$ of admissible states but with transition rate $\bar{\boldsymbol{q}}(\boldsymbol{m}+e_i,\boldsymbol{m}+e_j)$ for a change from $\boldsymbol{m}+e_i$ into $\boldsymbol{m}+e_j$ defined by:

For $i = 0,1,2$:

$$\begin{cases} \bar{\boldsymbol{q}}(\boldsymbol{m}+e_i,\boldsymbol{m}+e_{i+1}) = f_i(\boldsymbol{m}+e_i)b_{i+1}(\boldsymbol{m}) \\ \bar{\boldsymbol{q}}(\boldsymbol{m}+e_i,\boldsymbol{m}+e_{i-1}) = f_i(\boldsymbol{m}+e_i)b_{i+1}(\boldsymbol{m}) \end{cases} \tag{1.67}$$

Hence, for the exterior:

$$\begin{cases} \bar{\boldsymbol{q}}(\boldsymbol{m}+e_2,\boldsymbol{m}) = f_2(\boldsymbol{m}+e_2)b_0(\boldsymbol{m}) \\ \bar{\boldsymbol{q}}(\boldsymbol{m}+e_1,\boldsymbol{m}) = f_1(\boldsymbol{m}+e_1)b_2(\boldsymbol{m}) \\ \bar{\boldsymbol{q}}(\boldsymbol{m},\boldsymbol{m}+e_2) = \bar{\boldsymbol{q}}(\boldsymbol{m},\boldsymbol{m}+e_1) = \lambda b_1(\boldsymbol{m}) \end{cases} \tag{1.68}$$

In words that is, up to an exponential service scaling, the adjoint chain covers the original chain in natural flow direction but it also includes a proportional flow in opposite direction.

Result 1.5.7 *There exists a steady state solution $\pi(\boldsymbol{n})$ of (1.65) of the product form structure (with c a normalizing constant at* $\boldsymbol{C}$*):*

$$\pi(\boldsymbol{n}) = c\boldsymbol{H}(\boldsymbol{n}) \prod_{i=1,2} \left[\frac{1}{\mu_i}\right]^{n_i} \qquad (\boldsymbol{n} \in \boldsymbol{C}) \tag{1.69}$$

if and only if adjoint reversibility applies with solution $\boldsymbol{H}(\cdot)$. That is, for some function $\boldsymbol{H}(\boldsymbol{n})$ at $\boldsymbol{C}$*, the adjoint Markov chain is reversible, i.e. for any pair of states $\boldsymbol{n},\boldsymbol{n}' \in \boldsymbol{C}$:*

$$\boldsymbol{H}(\boldsymbol{n})\bar{\boldsymbol{q}}(\boldsymbol{n},\boldsymbol{n}') = \boldsymbol{H}(\boldsymbol{n}')\bar{\boldsymbol{q}}(\boldsymbol{n}',\boldsymbol{n}) \tag{1.70}$$

Proof. The proof is concluded directly by substitution of (1.67) in (1.66) or equivalently in (1.65) and showing equality for $(5.6.i) = (5.6.i)'$ for $i = 0,1,2$. □

Remark 1.5.8 (Adjoint reversibility) *Result 1.5.7 characterizes the existence of a product form solution by means of the so-called concept of reversibility, as will be defined below, of the adjoint Markov chain. Here we emphasize that the original system itself is* ***nót*** *reversible. The characterization will therefore be referred to as 'adjoint reversibility'.*

Remark 1.5.9 (Reversibility characterization) *The major advantage of result 1.5.7 is that it enables one to verify the existence of a product form of the form (1.69), by simply investigating the existence of a reversible solution $\boldsymbol{H}(\boldsymbol{n})$. This in turn, can be verified by the so-called Kolmogorov criterion (see [33]) as based upon just the transition rates as defined by (1.67). More precisely, either by checking whether for all cycles of transitions:*

$$\bar{\boldsymbol{q}}(\boldsymbol{n}_0,\boldsymbol{n}_1)\bar{\boldsymbol{q}}(\boldsymbol{n}_1,\boldsymbol{n}_2)\ldots\bar{\boldsymbol{q}}(\boldsymbol{n}_t,\boldsymbol{n}_0)=\bar{\boldsymbol{q}}(\boldsymbol{n}_0,\boldsymbol{n}_t)\bar{\boldsymbol{q}}(\boldsymbol{n}_t,\boldsymbol{n}_{t-1})\ldots\bar{\boldsymbol{q}}(\boldsymbol{n}_1,\boldsymbol{n}_0) \tag{1.71}$$

or, equivalently, whether for some fixed $\boldsymbol{n}_0 \in C$ and any state $\boldsymbol{n} \in \mathbf{C}$:

$$\boldsymbol{H}(\boldsymbol{n})=c\prod_{k=0}^{K-1}\left[\frac{\bar{\boldsymbol{q}}((\boldsymbol{n}_k\rightarrow\boldsymbol{n}_{k+1})}{\bar{\boldsymbol{q}}((\boldsymbol{n}_{k+1}\rightarrow\boldsymbol{n}_k)}\right] \quad \begin{array}{l}\textit{for any path } \boldsymbol{n}_0\rightarrow\boldsymbol{n}_1\rightarrow\ldots\rightarrow\boldsymbol{n}_K=\boldsymbol{n}\\ \textit{(for which the denominator is positive).}\end{array} \tag{1.72}$$

Remark 1.5.10 (Routing and service factorization) *Either of these 'adjoint reversibility' checks in turn can generally be reduced to basic cycles or short paths (also see section 1.5.4) that directly suggest a necessary form of $\boldsymbol{H}(\boldsymbol{n})$. This form in turn can generally be decomposed in a service and routing component, by*

$$\frac{\boldsymbol{H}(\boldsymbol{m}+e_i)}{\boldsymbol{H}(\boldsymbol{m}+e_j)}=\frac{x_i(\boldsymbol{m})}{x_j(\boldsymbol{m})}\frac{f_j(\boldsymbol{m}+e_j)}{f_i(\boldsymbol{m}+e_i)}=\boldsymbol{R}(\boldsymbol{m})\boldsymbol{S}(\boldsymbol{m}) \tag{1.73}$$

for some functions $\boldsymbol{R}(\boldsymbol{n})$ and $\boldsymbol{S}(\boldsymbol{n})$ provided, as for (1.66), both states $\boldsymbol{m}+e_i,\boldsymbol{m}+e_j \in$ $\boldsymbol{C}$. In addition, necessarily the numerator has to be equal to 0 if the denominator is equal to 0.

Here $\boldsymbol{R}(\boldsymbol{n})$ might be regarded as a component (solution) which only deals with the routing and thus also blocking, from one station to another, as determined by

$$\frac{\boldsymbol{R}(\boldsymbol{m}+e_i)}{\boldsymbol{R}(\boldsymbol{m}+e_j)}=\frac{\boldsymbol{x}_i(\boldsymbol{m})}{\boldsymbol{x}_j(\boldsymbol{m})} \text{ with } \{\boldsymbol{x}_i(\boldsymbol{m})\} \textit{ for any fixed underlying 'state' } \boldsymbol{m} \tag{1.74}$$

representing the local solutions of the local routing equations

$$\sum\nolimits_j \boldsymbol{x}_i(\boldsymbol{m})\bar{p}_{ij}(\boldsymbol{m})=\sum\nolimits_j \boldsymbol{x}_j(\boldsymbol{m})\bar{p}_{ji}(\boldsymbol{m}) \tag{1.75}$$

for the state dependent routing probabilities of the adjoint model $\bar{p_{ij}}(\boldsymbol{m})$ from $\boldsymbol{m}+e_i\rightarrow\boldsymbol{m}+e_j$ with $\boldsymbol{m}$ fixed. Similarly, $\boldsymbol{S}(\boldsymbol{n})$ represents the component (solution) for the service durations at the stations as by:

$$\frac{\boldsymbol{S}(\boldsymbol{m}+e_i)}{\boldsymbol{S}(\boldsymbol{m}+e_j)}=\frac{f_j(\boldsymbol{m}+e_j)}{f_i(\boldsymbol{m}+e_i)} \tag{1.76}$$

This factorization in a routing and service solution also seems to be characteristic for product form type expressions and can by itself be regarded as a 'product form' feature.

Remark 1.5.11 (Special examples) *In the subsequent sections 1.5.3-1.5.5 three types of examples will be provided to illustrate the possibility of product forms for tandem (or serial) structures despite the presence of station dependencies. These examples can be distinguished in examples with*

- *pure service dependence*
- *pure routing dependence*
- *or mixed*

1.5.3 Service examples

In this section assume that jobs upon arrival at the system or once having completed a service at a station can not be blocked, i.e. assume that

$$\begin{cases} \boldsymbol{R}(\cdot) \equiv 1 \\ b_0 \equiv b_1(\cdot) \equiv b_2(\cdot) = 1 \end{cases} \tag{1.77}$$

Example 1.5.12 (Independent services) *Clearly, the standard case with independent service capacities $f_i(n_i)$ at station i when n_i jobs are present is included by:*

$$\begin{cases} \boldsymbol{H}(n) = \boldsymbol{S}(n) = \lambda^{n_1+n_2} \prod_{i=1,2} \left[\prod_{k=1}^{n_i} f_i(k)\right]^{-1} \text{ and} \\ \bar{\boldsymbol{q}}(\boldsymbol{m}+e_i, \boldsymbol{m}+e_j) = f_i(n_i+1) \quad j = i+1, i-1 \quad (f_0(\cdot) = \lambda) \end{cases} \tag{1.78}$$

Example 1.5.13 (General function) *As a first situation with interdependence, often provided in the literature, suppose that for some functions strictly positive functions $\Phi(\cdot)$ and $\Psi(\cdot)$:*

$$f_i(\boldsymbol{m}+e_i) = \frac{\Phi(\boldsymbol{m})}{\Psi(\boldsymbol{m}+e_i)} \qquad (\textit{for all } \boldsymbol{m} \textit{ and } i) \tag{1.79}$$

Then one directly verifies (1.70), with $f_0(\mathbf{m}+e_0) = \lambda$, by

$$\frac{\boldsymbol{H}(\boldsymbol{m}+e_i)}{\boldsymbol{H}(\boldsymbol{m}+e_j)} = \frac{\boldsymbol{S}(\boldsymbol{m}+e_i)}{\boldsymbol{S}(\boldsymbol{m}+e_j)} = \frac{\bar{\boldsymbol{q}}(\boldsymbol{m}+e_j, \boldsymbol{m}+e_i)}{\bar{\boldsymbol{q}}(\boldsymbol{m}+e_i, \boldsymbol{m}+e_j)} =$$
$$\frac{f_j(\boldsymbol{m}+e_i)}{f_i(\boldsymbol{m}+e_j)} = \frac{\Psi(\boldsymbol{m}+e_i)}{\Psi(\boldsymbol{m}+e_j)}$$

by choosing

$$H(\boldsymbol{n}) = \lambda^{n_1+n_2}\Psi(\boldsymbol{n})$$

However, forms as for what type of functions $\Psi(\cdot)$ and $\Phi(\cdot)$ the service condition (1.79) is satisfied are not obvious. This is where the Kolmogorov criterion (1.71) or (1.72) might come in handy as will be illustrated in the next example.

Example 1.5.14 (Proportional and Unproportional Processor Sharing) *As an extension of standard processor sharing disciplines for one service location, in present-day service structures, such as Internet (cf. [4], [61]), a single service entity may have to share its capacity over multiple service stations, by*

$$f_i(n_1,n_2) = T(n_1+n_2)s_i(n_i \mid n_1+n_2)$$

where $T(\cdot)$ represents the total service capacity of the service entity and where $s_i(\cdot \mid \cdot)$ represents the fraction of this capacity allocated to station i. A processor sharing function by which each job present at any of the two stations (rather than standardly at each station separately) gets an equal (fair) share of the capacity, is hereby included by: $s_i(n_i \mid n_1+n_2) = n_i/(n_1+n_2)$. This would allocate capacity over both stations proportional to the workloads present. This will indeed still lead to a product form result as can be concluded directly from (1.76) or (1.79) or could also have been concluded indirectly from [10], with

$$\Psi(n) = \frac{1}{n_1!}\frac{1}{n_2!}\left[\prod_{k=1}^{n_1+n_2} T(k)\right]^{-1}$$

*But also **un**proportional sharing functions over both stations might still retain the necessary invariance (1.72) to secure a product form, for example*

$$s_i(n_i \mid n_1+n_2) = \begin{cases} \frac{2}{3} & ,i=1, \quad \frac{1}{3} \quad ,i=2, \quad n_1 > n_2 \\ \frac{1}{3} & ,i=1, \quad \frac{2}{3} \quad ,i=2, \quad n_1 < n_2 \\ \frac{1}{3} & ,i=1,2 \qquad\qquad\qquad n_1 = n_2 \end{cases} \tag{1.80}$$

In words that is, a double share is provided to the highest workload so as to strive for an equal workload at both stations (However, as a price to pay to satisfy the invariance condition (1.76) note that a capacity of $\frac{1}{3}$ is lost when $n_1 = n_2$). Condition (1.72) or (1.76) are now verified with

$$\boldsymbol{H}(\boldsymbol{n}) = \lambda^{n_1+n_2}\left[\prod_{k=1}^{n_1+n_2} T(k)\right]^{-1}\left[2^{\max(n_1,n_2)}\right]^{-1}[3]^{n_1+n_2}$$

This unproportional processor sharing product form possibility seems to be unreported and may lead to practical approximations.

1.5.4 Blocking examples

Throughout this subsection assume that the service capacities are independent and of the form $f_i(n_i)$ at station $i = 1,2$, with corresponding service solution $\boldsymbol{S}(n)$ as in example 1.5.12 in section 1.5.3. Hence,

$$\boldsymbol{H}(\boldsymbol{n}) = \boldsymbol{R}(\boldsymbol{n})\boldsymbol{S}(\boldsymbol{n}) \quad \text{with } S(\boldsymbol{n}) \quad \text{by (1.78) and} \tag{1.81}$$

$$\boldsymbol{R}(n) = 1_{\boldsymbol{C}}(n) \qquad \text{with } \boldsymbol{C} \text{ as specified below.} \tag{1.82}$$

General blocking condition.

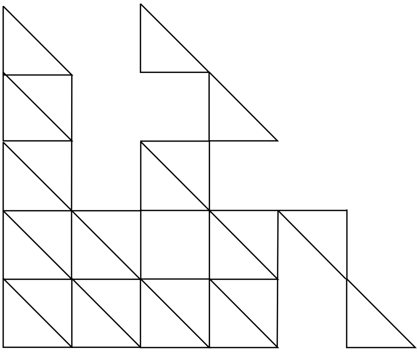


Fig. 1.20: Transition Structure.

In this two (dimensional) station case first observe (see figure 1.20) that any cycle of transitions as in criterion (1.71) can be seen as a regular structure that is built by two basic cycles of either form

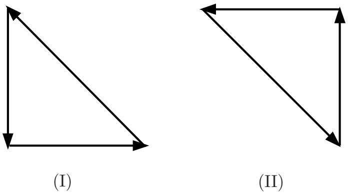


Fig. 1.21: Basic Cycles.

$$\begin{aligned} I \quad &: \ \boldsymbol{m} + e_1 \to \boldsymbol{m} + e_2 \to \boldsymbol{m} \to \boldsymbol{m} + e_1 \\ II \quad &: \ \boldsymbol{m} + e_1 \to \boldsymbol{m} + e_1 + e_2 \to \boldsymbol{m} + e_2 \to \boldsymbol{m} + e_1 \end{aligned}$$

The implicit assumption to be made here is that the adjoint transition rates within each of these basic cycles are consistently positive. That is, if a transition has a positive rate also its opposite one has a positive rate as according to the adjoint rates. This is a quite realistic assumption. (If it is not satisfied one certainly cannot expect a solution as of a form (1.69)).

By substituting the adjoint transition rates, leaving out the service rate functions $f_.(\cdot)$, the cycle condition (1.71) can now be applied to each of these two cycles. Cycle I then leads to the trivial condition:

$$b_2(\boldsymbol{m})b_0(\boldsymbol{m})b_1(\boldsymbol{m}) = b_1(\boldsymbol{m})b_2(\boldsymbol{m})b_0(\boldsymbol{m})$$

Cycle II however is verified if and only if

$$\begin{aligned} &b_1(\boldsymbol{m}+e_1)b_2(\boldsymbol{m}+e_2)b_0(\boldsymbol{m}) = \\ &b_2(\boldsymbol{m})b_1(\boldsymbol{m}+e_2)b_0(\boldsymbol{m}+e_1) \end{aligned} \tag{1.83}$$

As a consequence, condition (1.83) can thus be seen as a necessary and sufficient condition for adjoint reversibility to be satisfied with solution (1.69) and $\boldsymbol{H}(\cdot)$ as by (1.81) and (1.82). Let us provide some examples.

Example 1.5.15 (Finite capacity buffers) *As an extension of example 1.5.3 in section 1.5.1, suppose that both station 1 and station 2 have a finite capacity constraint for at most N_1 and N_2 (such as due to an intermediate buffer) jobs respectively. The equality condition (1.83) can then be verified for*

$$\begin{aligned} b_1(\boldsymbol{n}) &= 1_{(n_1<N_1, n_2<N_2)} && \textit{(block arrivals when either} \\ & && \textit{station 1 or station 2 is saturated)} \\ b_2(\boldsymbol{n}) &= 1_{(n_2<N_2)} && \textit{(stop station 1 if station 2 is saturated)} \\ b_0(\boldsymbol{n}) &= 1_{(n_1<N_1)} && \textit{(stop station 2 if station 2 is saturated)} \end{aligned}$$

(Note that the product form modification for example 1.5.3 is included when $n_2 = N_2$). For example, if $n_2+1 = N_2$ both $b_2(\mathbf{n}+e_2)$ in the left hand side and $b_1(\mathbf{n}+e_2)$ in the right hand side of (1.83) are equal to 0, and similarly if $n_1+1 = N_1$. Rather than just by an out = 0 ↔ in = 0 principle, as in the finite tandem example 1.5.3 in section 1.5.1, we have thus more formally proven the product form (1.81) with $\boldsymbol{S}(n)$ and $\boldsymbol{R}(n)$ as specified, with $\boldsymbol{C}$ the set of admissible states:

$$\boldsymbol{C} = \left\{ \boldsymbol{n} \;\middle|\; \begin{array}{l} n_1 \leq N_1 \\ n_2 \leq N_2 \\ n_1 + n_2 \neq N_1 + N_2 \end{array} \right\}$$

Remark 1.5.16 *In chapter 1.7 it will be numerically illustrated and formally proven that the product form result modification as in section 1.5.1 for example 1.5.3 provides simple bounds for the more natural finite tandem queue as merely specified by*

$$\begin{cases} b_1(\boldsymbol{n}) = 1_{(n_1<N_1)} \\ b_2(\boldsymbol{n}) = 1_{(n_2<N_2)} \end{cases}$$

As this natural finite tandem queue has no product form (as argued in example 1.5.3 and as it violates condition (1.83)), the characterization by adjoint reversibility, as leading to condition (1.83), can thus still be regarded as of practical interest.

More generally, example 1.5.15 can in fact be seen as a special case by setting $g_i(n_i) = 1_{(n_i<N_i)}$ when for given functions $g_1(n_1), g_2(n_2)$ at $\boldsymbol{C}$:

$$\begin{aligned} b_1(\boldsymbol{n}) &= g_1(n_1)g_2(n_2) \\ b_2(\boldsymbol{n}) &= g_2(n_2) \quad \textit{and} \\ b_0(\boldsymbol{n}) &= g_1(n_1) \end{aligned}$$

Condition (1.83) is satisfied and by (1.73) leads to

$$\begin{cases} \boldsymbol{R}(n) = \prod_{i=1}^{2} \left[\prod_{k=0}^{n_i-1} g_i(k)\right] \quad \text{at } \boldsymbol{C} \text{ with} \\ \boldsymbol{C} \text{ as in example 1.5.15 and} \\ N_i = \min\{k \mid g_i(k) = 0\} \end{cases} \tag{1.84}$$

Example 1.5.17 *Alternatively, condition (1.83) is also easily verified if for given functions $g(n), g_1(n_1), g_2(n_2)$:*

$$\begin{aligned} b_1(\boldsymbol{n}) &= d(n_1+n_2) \\ b_2(\boldsymbol{n}) &= d_1(n_1) \\ b_0(\boldsymbol{n}) &= d_2(n_2) \end{aligned}$$

with general solution

$$\boldsymbol{R}(\boldsymbol{n}) = \prod_{k=0}^{n_1+n_2} d(k) \prod_{i=1,2} \left[\prod_{k=M_i+1}^{n_i} d_i(k)\right]^{-1} \qquad \textit{at } \boldsymbol{C} \textit{ with}$$

$$\boldsymbol{C} = \left\{ \boldsymbol{n} \;\middle|\; \begin{array}{l} n_1 + n_2 \le M = \min\{k \mid g(k) = 0\} \\ n_i \ge M_i = \min\{k \mid g_i(k) = 0\} \\ i = 1,2 \end{array} \right\}$$

Total number blocking probability. For example, with $g_1(\cdot) = g_2(\cdot) \equiv 1$ but $d(k) = \alpha(k)$, arrivals can be assumed to be blocked with probability $\alpha(k)$ when $n = n_1 + n_2$ jobs are already present. Say when the arrival rate is thinned by a factor 2 for $n > M$ and completely blocked if $n = N > M$ we would obtain

$$\boldsymbol{R}(n) = 2^{-[n-M]^+} \qquad n \le N$$

Service delay - minimal workloads. Clearly, the functions $b_2(\cdot)$ and $b_0(\cdot)$ can now also be seen as delay factors $d_1(n_1)$, $d_2(n_2)$ and be included in the service rate functions $f_1(n)$ and $f_2(n)$ up to the point that these are assumed to be strictly positive as reflected by the expression for $\boldsymbol{S}(n)$. But also strict 0-values can now be included. As an example, with $d(\cdot) \equiv 1$ and

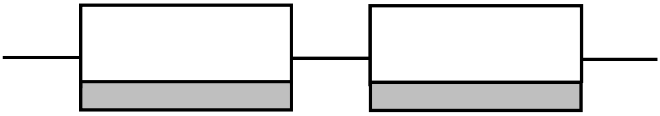


Fig. 1.22: Minimal Workloads.

$$d_i(n_i) = 1_{(n_i > N_i)} \qquad (i = 1,2)$$

we would block departures from and effectively stop the servicing of station i when it has reached a minimum of M_i jobs. In other words, the station should always have a minimum workload (which can also be regarded as safety buffers as in figure 1.22) of M_i jobs, as specified by the solution

$$\boldsymbol{R}(n) = 1_{(n_1 \geq M_1\,; n_2 \geq M_2)}$$

1.5.5 Mixed examples

Though in some of the earlier examples blocking at a station can also be reformulated as if the servicing (or arrival process) at the preceding (outside) station is stopped, let us give two more examples in which a mixed form is necessarily required to guarantee the adjoint reversibility condition.

In both examples the servicing is assumed to be processor sharing over both stations, say with a total service capacity $\Psi(n)$ when $n \leq n_1 + n_2$ jobs are present and only arrivals and servicing can be 'blocked' as to be parameterized by:

$$\begin{cases} f_1(n) = \Phi(n) s_1(n_1 \mid n_1 + n_2) \\ f_2(n) = \Phi(n) s_2(n_2 \mid n_1 + n_2) \\ b_1(\cdot) \equiv b_2(\cdot) \equiv 1 \end{cases}$$

Example 1.5.18 (Restricted state space) *Assume that the service sharing is proportional to the number of jobs present (similar to a standard processor sharing discipline), but that there is an inclination for keeping more jobs at station 1, by*

$$\begin{cases} s_1(n_1 \mid n_1+n_2) = [n_1/(n_1+n_2)]1_{(n_1 \geq n_2+1)} \\ s_2(n_2 \mid n_1+n_2) = [n_2/(n_1+n_2)] \\ b_0(\boldsymbol{n}) \qquad\qquad = 1_{(n_1 \geq n_2)} \end{cases}$$

with $\boldsymbol{C}$ *the restricted state space*

$$\boldsymbol{C} = \{n \mid n_1 \geq n_2 - 1 \geq 0\},$$

(1.70) is satisfied by

$$\boldsymbol{H}(\boldsymbol{n}) = \frac{1}{n_1!}\frac{1}{n_2!}\prod_{k=1}^{n}\Phi(k)$$

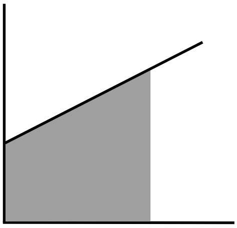


Fig. 1.23: Restricted state space.

Remark 1.5.19 *Though the solution looks standard, note that this example cannot be concluded by simply restricting the state space under reversibility conditions, as mentioned in [33] and [45], as the tandem queue itself is* ***not*** *reversible.*

Example 1.5.20 (Full service capacity) *Assume that there is just a single server whose capacity is fully devoted to one of the two stations (i.e.* $\Phi(k) = 1$*), as by*

$$\begin{cases} s_1(n_1 \mid n_1+n_2) = 1(n_1 = n_2+1 \vee n_1 = n_2+2) \\ s_2(n_2 \mid n_1+n_2) = 1(n_1 = n_2 \vee n_1 = n_2-1) \\ b_0(\boldsymbol{n}) \qquad\qquad = 1(n_1 = n_2 \vee n_1 = n_2+1) \end{cases}$$

The adjoint reversibility (1.70) is then satisfied at

$$\boldsymbol{C} = \{n \mid 0 \leq n_1 = n_1 - 1, n_2, n_2+1, n_2+2\}$$

as illustrated in figure 1.24 with $\boldsymbol{H}(\boldsymbol{n}) = [n_1 n_2]^{-1}$

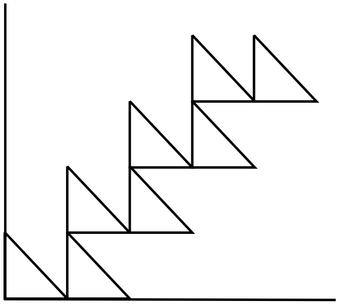


Fig. 1.24: Full service example.

1.6 Jacksonian clusters

As shown by example 1.5.3, even for the most simple case of a two station open network one cannot generally expect a product form when restricted capacities become involved to accommodate jobs (e.g. by finite buffers). Nevertheless, under specific, more unnatural protocols and possibly enforced by modification, specific product form results could still be concluded.

Such results can still be of practical interest such as to provide reasonable orders of magnitude or bounds. However, can these specific product form results also be expected for larger networks, such as most standardly an arbitrary Jackson type network with finite capacities?

In this section, it will be shown analytically by just two specific applications that the results as in section 1.5 for a 'simple' finite tandem queue can indeed also be extended to restricted Jackson type networks.

In the next section, as of more practical interest, it will merely be argued and be illustrated numerically how the concept of station balance and the product form results from section 1.5 can also be extended to provide practical numerical results for assembly type networks with restricted Jacksonian clusters. First in section 1.6.1 let us briefly review the notion and product form result of a standard Jackson network. Here, for its illustrative purpose without restriction of generality and for its broader use in section 1.7, we only consider an open Jackson network.

1.6.1 A Jackson cluster

Consider an open network with J service stations, numbered $1,\dots,J$ and Poisson arrival rate with parameter γ_j at station $j = 1,\dots,J$. After a service completion at station i a job will instantaneously

- route to a next service station j with probability p_{ij} $(j \neq 0)$, or
- leave the system with probability $p_{i0} = [1 - \sum_{j \neq i} p_{ij}]$.

At each visit at station i a job requires an exponential amount of service with parameter μ_i. Station i services at a service capacity $f_i(n_i)$ when n_i jobs are present. Let the vector

$$\boldsymbol{n} = (n_1, \ldots, n_J)$$

denote the number of jobs n_i at station $i = 1, \ldots, J$ and let $\boldsymbol{e}_i$ be the unit vector for component i. Hence, $\boldsymbol{n} + e_i$ and $\boldsymbol{n} - e_i$ denote the vectors equal to n with one job more respectively less at station i and $\boldsymbol{n} - e_i + e_j$ indicates that one job has moved from station i to j. Furthermore, again use the notational convention that $\boldsymbol{n} + \boldsymbol{e}_0 = \boldsymbol{n} - e_0 = \boldsymbol{n}$, write $p_{0j} = \gamma_j / \lambda$ with $\lambda = \sum_j \gamma_j$ and let

$$F(\boldsymbol{n}) = \prod_{i=1} \left[\prod_{k=1}^{n_i} f_i(k) \right]^{-1}$$

For the unrestricted case, that is with unlimited state space
$\boldsymbol{C}_\infty = \{\boldsymbol{n} \mid n_i \geq 0 \;,\; i = 1, \ldots, J\}$ the global balance relations

$$\begin{array}{ll}
\left\{\begin{array}{l} \pi(\boldsymbol{n})\lambda + \\ \pi(\boldsymbol{n}) \sum_j \mu_j f_j(n_j) \end{array}\right\} & \begin{array}{l} (1.85.0) \\ (1.85.j) \end{array} \\
= & \\
\left\{\begin{array}{l} \displaystyle\sum_i \pi(\boldsymbol{n} + e_i) \mu_i f_i(n_i + 1) p_{i0} + \\ \displaystyle\sum_j 1_{(n_j > 0)} [\pi(\boldsymbol{n} - e_j)\gamma_j + \sum_i \pi(\boldsymbol{n} + e_i - e_j) \mu_i f_i(n_i + 1) p_{ij}] \end{array}\right\} & \begin{array}{l} (1.85.0)' \\ (1.85.j)' \end{array}
\end{array} \qquad (1.85)$$

are then directly verified (see remark 1.6.3 for its detail) by the station balance relations $(1.85.j) = (1.85.j)'$, for each station $j = 0, 1, 2, \ldots, J$ separately, i.e. for any station $j \neq 0$ and state $\boldsymbol{n}$ with $n_j > 0$ by

$$\pi(\boldsymbol{n}) \mu_j f_j(n_j) = \sum_i \pi(\boldsymbol{n} + e_i - e_j) \mu_i p_{ij} + \pi(\boldsymbol{n} - e_j)\gamma_j \tag{1.86}$$

and for the exterior (station $j = 0$) and each state $\boldsymbol{n}$:

$$\pi(\boldsymbol{n})\lambda = \sum_i \pi(\boldsymbol{n} + e_i) \mu_i f_i(n_i + 1) p_{i0} \tag{1.87}$$

by assuming the product form

$$\pi(\boldsymbol{n}) = cF(\boldsymbol{n}) \prod_i \left[\frac{\lambda_i}{\mu_i} \right]^{n_i} \qquad \text{where} \tag{1.88}$$

$$\lambda_j = \gamma_j + \sum_i \lambda_i p_{ij} \qquad (j = 1, \ldots, J) \tag{1.89}$$

Remark 1.6.1 (Traffic equations) *Here the implicit natural assumption is made that the so-called traffic equations (1.89) have a unique solution $\{\lambda_j\}$.*

Remark 1.6.2 (Decomposability) *In fact, one might note that for the present unrestricted case we can also factorize the normalizing constant c and hence the steady state solution $\pi(n)$ as if the stations can be regarded as being independent.*

Remark 1.6.3 (Verification of (1.85)) *To verify (1.86) and (1.87), by assuming (1.88) we can substitute*

$$\begin{aligned}
\left[\frac{\pi(\boldsymbol{n}+\boldsymbol{e}_i-\boldsymbol{n}_j)}{\pi(\boldsymbol{n})}\right] &= \left[\frac{f_j(n_j)}{f_i(n_i+1)}\right]\left[\frac{\lambda_i}{\lambda_j}\right]\left[\frac{\mu_j}{\mu_i}\right]\\
\left[\frac{\pi(\boldsymbol{n}+\boldsymbol{n}_i)}{\pi(\boldsymbol{n})}\right] &= \left[\frac{1}{f_i(n_i+1)}\right]\left[\frac{\lambda_i}{\mu_i}\right]\\
\left[\frac{\pi(\boldsymbol{n}-\boldsymbol{n}_j)}{\pi(\boldsymbol{n})}\right] &= \left[\frac{\mu_j}{\lambda_j}\right][f_j(n_j)]
\end{aligned}$$

By dividing by $\pi(\boldsymbol{n})$ and cancelling terms, (1.86) for $j \neq 0$ with $n_j > 0$ then reduces to the traffic equations (1.89). Similarly (1.87) is verified by also using (recalling (1.89) again):

$$\textstyle\sum_i \lambda_i p_{i0} = \sum_i \lambda_i[1-\sum_j p_{ij}] = \sum_j \lambda_j - \sum_i \lambda_i p_{ij} = \sum_i \gamma_j \qquad (1.90)$$

1.6.2 A restricted Jackson cluster

As a first direct extension now assume that a Jackson cluster of section 1.6.1 is constrained by:

- no more than N jobs in total and
- no less than M jobs in minimum

> **Loss and recycle protocol:**
> *If upon arrival N jobs are already present, an arriving job is blocked and lost. Conversely, if upon system departure the number of jobs left behind would drop below M, the departing job is recycled into the system at station j with probability $p_{0j} = \gamma_j / \sum_i \gamma_i$.*

Clearly, in order for a steady state solution to exist, the system has to be initiated in a state with $n \geq M$. In that case, the set of admissible states is restricted to:

$$\boldsymbol{S} = \{\boldsymbol{n} \mid n_i \geq 0\ ,\ i = 1,\ldots,M\ ;\ M \leq n \leq N\}$$

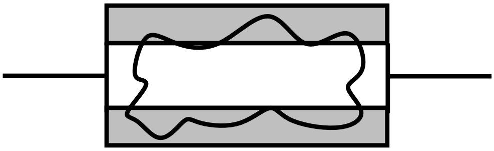


Fig. 1.25: Maximal and Minimal Workloads.

Result 1.6.4 (Recycle protocol) *Under the loss and recycle protocol the product form (1.88) remains valid restricted to* $\boldsymbol{C}$.

Proof. The upper limit N is directly taken into account by (1.87) as

$$\pi(\boldsymbol{n})\lambda 1_{(\boldsymbol{n}<N)} = \sum_i 1_{(\boldsymbol{n}<N)}\pi(\boldsymbol{n}+\boldsymbol{e}_i)\mu_i f_i(n_i+1)p_{i0} \tag{1.91}$$

is verified as before for $n < N$, (see remark 1.6.3) while for $n = N$ both sides of (1.91) are equal to 0.

Conversely, in order to take the lower limit M into account, for any $\boldsymbol{n} \in \boldsymbol{C}$, hence with $\boldsymbol{n} \le N$, the station balance relation (1.86) is to be replaced by

$$\begin{aligned}&\pi(\boldsymbol{n})\mu_j f_j(n_j) = \\ &\textstyle\sum_i \pi(\boldsymbol{n})\mu_i p_{ij} + 1_{(n_j>0)} \cdot \\ &\left[1_{(\boldsymbol{n}>M)}\pi(\boldsymbol{n}-\boldsymbol{e}_j)\gamma_j + 1_{(\boldsymbol{n}=M)}\pi(\boldsymbol{n}+\boldsymbol{e}_i-\boldsymbol{e}_j)\mu_i p_{i0} f_i(n_i+1)p_{0j}\right]\end{aligned} \tag{1.92}$$

Clearly, for $\boldsymbol{n} > M$, this is verified as for (1.86) by substituting (1.88). For $\boldsymbol{n} = M$, however, after substituting (1.88) and cancelling terms, again we need to use (1.90). □

To some extent the 'recycle protocol' as described above can be regarded as most natural as it allows services to continue. It only requires blocked departures to be reserviced. Alternatively, a seemingly stronger stop protocol could also be thought of stated as:

Stop protocol:

Stop the servicing of all stations if a departure from the system (Jackson cluster) is not allowed, in this case if $n = M$.

In the next section we will also consider multiple clusters of Jackson networks in which a departure from one cluster can be blocked due to a finite constraint at a next cluster. Purely for the purpose of providing a simple product form bound the somewhat simpler 'stop protocol' will then be more appropriate as the two protocols generally lead to exactly the same product form. This is shown below for the present situation of a guaranteed minimal workload.

Result 1.6.5 (Stop protocol) *Under the stop protocol the product form (1.88) remains valid restricted to* $\boldsymbol{C}$.

Proof. Again, the upper limit is directly taken into account as by (1.91) to replace (1.87). As for the lower limit constraint M, also the verification of the station balance (1.86), as given by (1.92), the recycle protocol, even becomes more direct for the stop protocol by

$$\begin{aligned}&\pi(\boldsymbol{n})\mu_j f_j(n_j)1_{(\boldsymbol{n}>M)} = \\ &\pi(\boldsymbol{n}-\boldsymbol{e}_j)\gamma_j 1_{(\boldsymbol{n}>M)} + \textstyle\sum_i \pi(\boldsymbol{n}+\boldsymbol{e}_i-\boldsymbol{e}_j)\mu_i f_i(n_i+1)p_{ij}1_{(\boldsymbol{n}>M)}\end{aligned} \tag{1.93}$$

where it is noted again that the state $\boldsymbol{n}-\boldsymbol{e}_j$ is not admissible when $\boldsymbol{n}=M$, which directly reduces to (1.86) with $1_{(\boldsymbol{n}>M)}$ at both hand sides. □

1.6.3 A conservative product form protocol

Now consider a Jackson network as described in section 1.6.1 but with a finite capacity constraint for no more than N_j jobs at station j; $j=1,2,\dots,J$. (Here one or more of the values N_j can be infinite). Clearly, as already shown by the tandem case in section 1.5.1, under a natural blocking protocol by which an upstream station is blocked when a next downstream station is congested a product form cannot be expected as station balance is necessarily violated. In line with the product form modification in section 1.5.1 and example 1.5.15 for the finite tandem example, however, a product form can be expected under the, as it is called here:

> **Conservative protocol.**
> *When a station* j *is congested, i.e.* $n_j = N_j$
> *stop all other stations* $l \neq j$ *and stop arrivals.*

Result 1.6.6 (Conservative protocol) *Under the conservative protocol, the product (1.88) remains valid restricted to*

$$\boldsymbol{C} = \left\{\boldsymbol{n} \mid 0 \le n_i \le N_i\,;\ i=1,2,\dots,J\,;\ n_i+n_j < N_i+N_j \mathit{\ for\ all\ pairs\ \ } i\neq j\right\}$$

Proof. Again we will verify the station balance relation (1.86) for $j\neq 0$ and $j=0$ in its present adapted form. Consider a fixed state $\boldsymbol{n}\in\boldsymbol{C}$. Then (1.86) for station $j=0$ (also for $n_j=N_j$) is to be replaced by:

$$\begin{aligned}&\pi(\boldsymbol{n})\mu_j f_j(n_j)\left[\textstyle\prod_{l\neq j} 1_{(n_l<N_l)}\right] = \\ &\pi(\boldsymbol{n}-\boldsymbol{e}_j)\gamma_j\left[\textstyle\prod_{l\neq j} 1_{(n_l<N_l)}\right] + \\ &\textstyle\sum_i \pi(\boldsymbol{n}+\boldsymbol{e}_i-\boldsymbol{e}_j)1_{(n_i+1\le N_i)}\left[\textstyle\prod_{l\neq i,j} 1_{(n_l<N_l)}\right]\mu_i f_i(n_i+1)p_{ij}\end{aligned} \tag{1.94}$$

As a state $\boldsymbol{n}+\boldsymbol{e}_j-\boldsymbol{e}_j$ can only be admissible if $n_i+1 \leq N_i$. Hence, as $1_{(n_i+1<Ni)} = 1_{(n_i<N_i)}$, either all terms in both hand sides are equal to 0 if $n_l = N_l$ for some $l \neq j$, or all indicator functions are equal to 1 so that (1.94) is identical to (1.86), as satisfied by the product form (1.88). Similarly, for $j=0$, the total outrate and inrate for the system are equated by (1.88) at $\boldsymbol{C}$ as:

$$\pi(\boldsymbol{n})\lambda\left[\prod_l 1_{(n_l<N_l)}\right] = \sum\nolimits_i \pi(\boldsymbol{n}+\boldsymbol{e}_i)1_{(n_i+1\leq N_i)}\left[\prod_l 1_{(n_l<N_l)}\right]\mu_i f_i(n_i+1)p_{i0} \quad (1.95)$$

□

Remark 1.6.7 (Conservative protocol) *The conservative protocol is referred to as conservative as it only continues the service at that station which resolves the congestions. As a consequence it avoids that more than one station can become congested at the same time.*

Remark 1.6.8 (Jump-over protocol) *A(nother) protocol to generally ensure the product form (1.88) at*

$$\boldsymbol{C} = \{\boldsymbol{n} \mid 0 \leq n_i \leq N_i\ ;\ i=1,\ldots,J\}$$

is by the

> ***Jump-over protocol.***
>
> *Let jobs jump over a saturated station i with $n_i = N_i$ to a next service station j according to the routing probabilities p_{ij}.*

The product form can be argued intuitively by assuming infinite capacities but a service speed $f_i(N_i+1) \to \infty$, for all i, so that the probability for a state with more than N_i jobs at any station i becomes virtually 0. An analytic proof, as based upon absorbing Markov chains, can be found in [52].

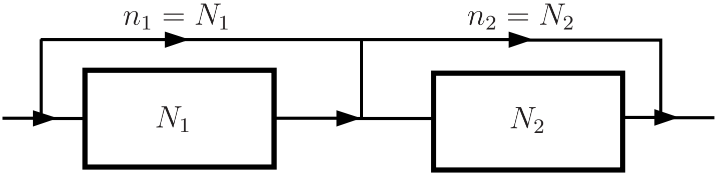


Fig. 1.26: Tandem Queue with Jump-Over.

As an illustrative example, though, as of practical interest by itself and in line with section 1.5, let us just reconsider the finite tandem queue with capacity constraints N_1 and N_2 at stations 1 and 2 jump-over protocol; i.e.

If $n_2 = N_2$ departures from station 1 clear the system.
If $n_1 = N_1$ arrivals are directly routed to station 2.
(If both $n_1 = N_1$ and $n_2 = N_2$ the arrivals are lost)

The global balance relations now become

$$\left\{\begin{array}{l} \pi(\boldsymbol{n})\mu_1 f_1(n_1)+ \\ \pi(\boldsymbol{n})\mu_2 f_2(n_2)+ \\ \pi(\boldsymbol{n})\lambda\left[1_{(n_1<N_1)} + 1_{(n_1=N_1)}1_{(n_2<N_2)}\right] \end{array}\right\}$$
$$=$$
$$\left\{\begin{array}{l} \pi(n)\lambda+ \\ \pi(\boldsymbol{n}+\boldsymbol{e}_1-\boldsymbol{e}_2)\mu_1 f_1(n_1+1)1_{(n_1<N_1)} + \pi(\boldsymbol{n}-\boldsymbol{e}_2)\lambda 1_{(n_1=N_1)}+ \\ \pi(\boldsymbol{n}+\boldsymbol{e}_2)\mu_2 f_2(n_2+1)1_{(n_2<N_2)} + \pi(\boldsymbol{n}+\boldsymbol{e}_1)\mu_1 f_1(n_1+1)1_{(n_2=N_2)}1_{(n_1<N_1)} \end{array}\right\} \tag{1.96}$$

By noting that for any state $\boldsymbol{n} \in \boldsymbol{C}$ with $(n_1,n_2) \neq (N_1,N_2)$:

$$1_{(n_1<N_1)} + 1_{(n_2=N_2)}1_{(n_2<N_2)} = 1_{(n_2<N_2)} + 1_{(n_2=N_2)}1_{(n_1<N_1)} = 1,$$

again these in turn are verified by a 'station balance' relation for each station $j = 0,1,2$ separately when substituting the product form:

$$\pi(\boldsymbol{n}) = cF(\boldsymbol{n})\left[\frac{\lambda}{\mu_1}\right]^{n_1}\left[\frac{\lambda}{\mu_2}\right]^{n_2} \qquad (n_1 \leq N_1\ ;\ n_2 \leq N_2) \tag{1.97}$$

Remark 1.6.9 *The 'tandem example' provided above is of some natural interest for present day packet switch communication structures (such as internet) in which case a load congestion might be skipped which will lead to only a partial loss of packets (information).*

1.7 Product form bounds for networks of restricted clusters

As mentioned before and shown by example 1.5.2 in section 1.5.1, a simple tandem queue with a finite capacity constraint already violates station balance and hence a product form. Nevertheless, as shown by its modification in example 1.5.3, its extension in example 1.5.15. for a tandem with two finite stations, and in section 1.6 for Jacksonian type clusters, under appropriate blocking protocols a product form expression might still be obtainable, possibly enforced by modification.

These product forms in turn, even though the protocols might be 'unnatural', might still be useful for the original non-product form system to provide a reasonable approximation or bound, as announced in remark 1.5.15. More precisely, for the simple but unsolvable tandem queue with both a finite first and a finite second station (as in example 1.5.2), the product form (modification) as in example 1.5.15 turns out to be quite fruitful to provide a simple (lower and upper) product-form bound for the loss probability (and throughput). As some numerical results for this two-station example can also be found in the chapter on error bounds and comparison results (chapter 7), in section 1.7.1 below some numerical support will directly be presented for a slightly larger four station tandem example.

In fact, in practical situations capacity constraints are often imposed upon clusters (groups) of stations rather than individual stations. In this section, therefore, it will merely be illustrated, in line with the two station tandem example and the results for a single cluster as in section 1.6, how the product-form modification approach also extends to and can be fruitful for larger networks, particularly 'assembly line or tandem type' structures with restricted Jacksonian clusters. Roughly speaking, this bounding approach is based on the two concepts of:

(i) regarding a cluster of stations with some common capacity constraint as 'one aggregate station'.

(ii) a modification of the system such that both the notion of station balance for individual stations, and of station balance for 'aggregate stations': referred to as 'cluster balance', are restored and satisfied.

1.7.1 Instructive tandem extension

In production environments, capacity constraints are often imposed upon clusters of workstations rather than individual workstations. It would thus be appealing if the principle of station balance can also be extended to a cluster level, by regarding a cluster as one aggregated station.

Consider, for example, the cluster extension of the tandem case (see figure 1.27) with four stations to be seen as a two-cluster model with capacity constraints T_1 and T_2 for the total number of jobs in cluster 1 (stations 1 and 2) and cluster 2 (stations 3 and 4).

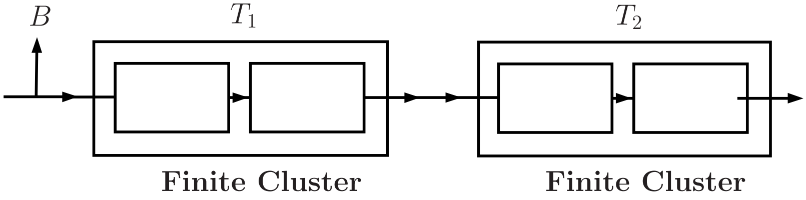


Fig. 1.27: Finite Cluster Extension

In order to enforce a simple product-form expression, a similar modification as example 1.5.3 in section 1.5.1 and example 1.5.15 in section 1.5.4 would then seem appealing at cluster level by also rejecting jobs at cluster 1 when cluster 2 is congested. This would lead to an upper bound for the blocking probability. In other words, at first glance we would expect a similar simple product-form bounding approach by simply regarding a cluster as a station and transforming the notion of balance per station into balance per cluster by just keeping track of the total number of jobs at each cluster. This will be referred to as *cluster balance*.

Example 1.7.1 *To be more precise, consider the simple assembly line structure with 4 service stations, numbered* $1,\dots,4$ *and finite capacity constraints* T_1 *for the total number of jobs at stations 1 and 2 (cluster 1) and* T_2 *at stations 3 and 4 (cluster 2). The system has an arrival rate of* λ *jobs per unit of time and assume that station* i *has (an exponential) service rate* $\mu_i f_i(k)$ *when* k *jobs are present. As before, let* n_i *denote the number of jobs at station* i*,* $i = 1,\dots,4$ *and* t_j *the total number of jobs at cluster* j*,* $j = 1,2$*. (*$t_1 = n_1 + n_2$ *and* $t_2 = n_3 + n_4$*). When the first cluster is saturated (*$t_1 = T_1$*) an arriving job is lost. When the second cluster is saturated (*$t_2 = T_2$*) the service at cluster 1 (that is at both stations) is stopped. As simple as the system may look to analyze, there is no simple expression for the loss probability* $\boldsymbol{B}$ *of arriving jobs or the throughput* $\boldsymbol{H} = \lambda(1-\boldsymbol{B})$.

In this example, both the notion of balance per station (as by (1.3) in section 1.2.1) and of balance per cluster (that is, as if a cluster is regarded as one aggregated station) are violated, since when $t_1 < T_1$ but $t_2 = T_2$:

- *the out-rate of stations 1 and 2 and the out-rate of cluster 1 are necessarily equal to 0 while the in-rate for station 1 (and possibly also for 2) and for cluster 1 are positive.*

The following artificial modification to enforce these notions can therefore be suggested.

- *When cluster 2 is saturated* ($t_2 = T_2$)*: stop the input.*
- *When cluster 1 is saturated* ($t_1 = T_1$)*: stop cluster 2 (that is, both stations at cluster 2).*

Indeed, under this modification one easily verifies the global balance (1.98) by station balance equations $(1.98.i) = (1.98.i)'$ for $i = 1,\dots,5$ at S_U the set of admissible states:

$$\boldsymbol{S}_U = \{\boldsymbol{n} \mid t_1 = n_1 + n_2 \le T_1 \; ; \; t_2 = n_3 + n_4 \le T_2 \; ; \; t_1 + t_2 \ne T_1 + T_2\}$$

as

$$\left\{\begin{array}{l} \pi(\boldsymbol{n})\mu_1(n_1)f_1(n_1)1_{(t_2<T_2)}+ \\ \pi(\boldsymbol{n})\mu_2 f_2(n_2)1_{(t_2<T_2)}+ \\ \pi(\boldsymbol{n})\mu_3 f_3(n_3)1_{(t_1<T_1)}+ \\ \pi(\boldsymbol{n})\mu_4 f_4(n_4)1_{(t_1<T_1)}+ \\ \pi(\boldsymbol{n})\lambda 1_{(t_1<T_1)}1_{(t_2<T_2)} \end{array}\right\} \begin{array}{l} (1.98.1) \\ (1.98.2) \\ (1.98.3) \\ (1.98.4) \\ (1.98.5) \end{array}$$

$$= \qquad (1.98)$$

$$\left\{\begin{array}{l} \pi(\boldsymbol{n}-e_1)1_{(n_1>0)}\lambda 1_{(t_2<T_2)}+ \\ \pi(\boldsymbol{n}-e_2+e_1)1_{(n_2>0)}\mu_1 f_1(n_1+1)1_{(t_2<T_2)}+ \\ \pi(\boldsymbol{n}-e_3+e_2)1_{(n_3>0)}\mu_2 f_2(n_2+1)1_{(t_1<T_1)}+ \\ \pi(\boldsymbol{n}-e_4+e_3)1_{(n_4>0)}\mu_3 f_3(n_3+1)1_{(t_1<T_1)}+ \\ \pi(\boldsymbol{n}+e_4)\mu_4 f_4(n_4+1)1_{(t_1<T_1)}1_{(t_2<T_2)} \end{array}\right\} \begin{array}{l} (1.98.1)' \\ (1.98.2)' \\ (1.98.3)' \\ (1.98.4)' \\ (1.98.5)' \end{array}$$

by substituting the product-form

$$\pi(\boldsymbol{n}) = c\lambda^{n_1+n_2+n_3+n_4}\prod_{i=1}^{4}\left\{\mu_i^{n_i}\left[\prod_{k=1}^{n_i} f_i(k)\right]\right\}^{-1} \quad , \boldsymbol{n} \in \boldsymbol{S}_U \qquad (1.99)$$

with c a normalizing constant. Clearly, the modification leads to an upper bound $\boldsymbol{B}_U \geq \boldsymbol{B}$ for the loss probability

$$\boldsymbol{B}_U = \sum_{\{\boldsymbol{n}\,|\,t_1=T_1 \text{ or } t_2=T_2\}} \pi_U(\boldsymbol{n}) \qquad (1.100)$$

Conversely, also a lower bound product-form modification $\boldsymbol{B}_L$ can be suggested by only rejecting arriving jobs when the total number of jobs $n_1+n_2+n_3+n_4 = T_1+T_2$ and allowing up to this number to be present at any station. Then (1.99) applies with $\boldsymbol{S}_U$ replaced by:

$$\boldsymbol{S}_L = \{\boldsymbol{n} \mid n_1+n_2+n_3+n_4 \leq T_1+T_2 \; ; \; n_i \geq 0 \; , \; i = 1,\ldots,4\} \qquad (1.101)$$

Below some numerical results are given for the case of single server stations. Here μ_i represents the service speed of station i, $\boldsymbol{B}_L$ and $\boldsymbol{B}_U$ are the easily obtained lower and upper bound for the blocking probability, $\boldsymbol{B}_{av} = (\boldsymbol{B}_L+\boldsymbol{B}_U)/2$ and $\boldsymbol{B}$ is obtained by numerical computation.

Remark 1.7.2 (Insensitive bounds) *Referring to sections 1.2.5 and 1.3.3, recall that for pure multi-server or processor sharing disciplines, the product-form expression (1.99) remains valid for arbitrary service distributions with means $1/\mu_i$. Also the bounds $\boldsymbol{B}_U$ and $\boldsymbol{B}_L$ can then be expected to be insensitive.*

Table 1.2: Lower and upper bounds of the loss probability $\boldsymbol{B}$ (and throughput $\boldsymbol{H}$ by $\boldsymbol{H} = \lambda(1-B)$) for finite two-cluster tandem example.

μ_1	μ_2	μ_3	μ_4	$\boldsymbol{T}_1$	$\boldsymbol{T}_2$	$\boldsymbol{B}_L$	$\boldsymbol{B}_U$	$\boldsymbol{B}_{av}$	$\boldsymbol{B}$
1	1	1	1	3	5	.33	.52	.43	.42
1	1	1	1	6	6	.25	.40	.33	.30
1	1	1	1	8	8	.20	.33	.27	.24
2	2	1	1	10	10	.10	.17	.14	.12
1	2	3	2	10	10	.054	.101	.078	.084
1.1	2	3	2	10	10	.021	.065	.048	.049

1.7.2 A Jackson Tandem

The simple two-station tandem clusters can directly be replaced by a Jackson cluster as from section 1.6 by applying either of the two protocols from section 1.6, the recycle or stop protocol for the entire cluster, if its departures are to be blocked.

- *The modification at cluster level as above if either the first or second cluster is congested*

More concrete, consider the situation of two finite Jackson clusters labeled $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$. As before let t_1 and t_2 be the total number of jobs at cluster 1 and at cluster 2.

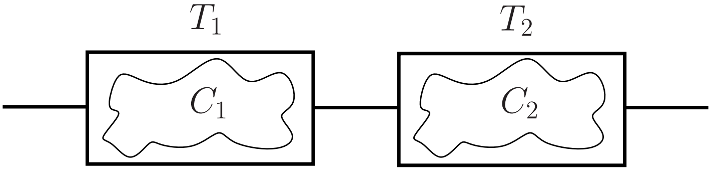


Fig. 1.28: Finite Jacksonian Tandem.

For the original system of interest assume that when a departure from cluster 1 say from station i is blocked, it has to undergo a new service at station i. Effectively, this means that as in the example above that only the departure station 2 is delayed or completely stopped when a congestion takes place. By the modification:

PF-modification

- *When a cluster is activated: stop arrivals and all stations at the other cluster*

a product form can now be expected. Indeed, for any station $j \in \boldsymbol{C}_1$ and with $n_j > 0$, the station balance then becomes:

$$\begin{aligned}&\pi(\mathbf{n})\mu_j f_j(n_j)1_{(t_2<T_2)} = \\ &\pi(\mathbf{n}-e_j)\gamma_j 1_{(t_2<T_2)} + \textstyle\sum_{i\in\mathbf{C}_1} 1_{(t_2<T_2)}\pi(\mathbf{n}+e_i-e_j)\mu_i f_i(n_1+1)p_{ij}\end{aligned} \qquad (1.102)$$

and for $j \in \mathbf{C}_2$

$$\begin{aligned}&\pi(\mathbf{n})\mu_j f_j(n_j)1_{(t_1<T_1)} = \\ &1_{(t_1<T_1)}\textstyle\sum_{i\in\mathbf{C}_1}\pi(\mathbf{n}+e_i-e_j)\mu_i f_i(n_i+1)p_{ij}+ \\ &1_{(t_1<T_1)}\textstyle\sum_{i\in\mathbf{C}_2}\pi(\mathbf{n}+e_i-e_j)\mu_i f_i(n_i+1)p_{ij}\end{aligned} \qquad (1.103)$$

Here in the first term in the r.h.s. of (1.103) it is noted that any state $n+e_i-e_j$ necessarily has $t_2-1<T_2$ jobs at cluster 2 so that the servicing at the first cluster is not stopped. Finally, the outrate and inrate equation for the system (station 0) are:

$$\begin{aligned}&\pi(\mathbf{n})\lambda 1_{(t_1<T_1)}1_{(t_2<T_2)} = \\ &\textstyle\sum_{i\in\mathbf{C}_2} 1_{(t_2<T_2)}\pi(\mathbf{n}+e_i)\mu_i f_i(n_1+1)1_{(t_1<T_1)}p_{i0} = \\ &1_{(t_1<T_1)}1_{(t_2<T_2)}\textstyle\sum_i \pi(\mathbf{n}+e_i)\mu_i f_i(n_1+1)p_{i0}\end{aligned} \qquad (1.104)$$

as $p_{i0}=0$ for $i\in\mathbf{C}_1$. By cancelling the equal indicator terms $1_{(t_1<T_1)}$ and $1_{(t_2<T_2)}$ in the left hand and right hand sides, each of these relations (1.102), (1.103) and(1.104), which together form the global balance equations, are now verified directly as before in section 1.6.1 by using the traffic relations (1.89) and (1.90) as for a standard Jackson cluster when substituting the product form:

$$\pi(\mathbf{n}) = c\prod_i \left[\frac{\lambda_i}{\mu_i}\right]^{n_i}\left[\prod_{k=1}^{n_i} f_i(k)\right]^{-1} \qquad (1.105)$$

at

$$\mathbf{C} = \left\{\mathbf{n} \mid t_1 = \textstyle\sum_{i\in\mathbf{C}_1} n_i \le T_1 \; ; \; t_2 = \textstyle\sum_{i\in\mathbf{C}_2} n_i \le T_2 \; ; \; t_1+t_2 \ne T_1+T_2\right\}$$

Remark 1.7.3 (Recycle protocol) *In line with section 1.6.2, the same product form (1.105) result can also be proven if, for either of the two Jackson clusters or for both, a recycle protocol would be applied upon departure blocking. (I.e. a departing job would be recycled into that cluster as a newly arriving job - here for cluster 2 it would then have to be assumed that $p_{ij}=\beta_j$ for $j\in C_2$ and all $i\in\mathbf{C}_1$).*

Remark 1.7.4 (Jump-over protocol) *In line with the jump-over protocol in remark 1.6.8, a(nother) product form (modification) protocol would be to let arriving jobs jump-over cluster 1 if $t_1=T_1$ and let jobs leaving cluster 1 clear the system if $t_2=T_2$. In that case (1.105) would apply with* ***C*** *restricted to*

$$\mathbf{C} = \{n \mid t_1 \le T_1 \; ; \; t_2 \le T_2\}$$

1.7.3 A nested case

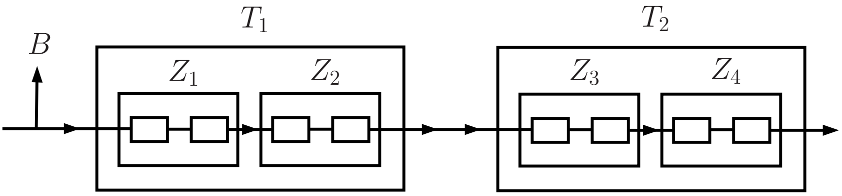


Fig. 1.29: Nested Finite Constraints.

A nested extension of the example from section 1.7.1, now consider two clusters in tandem each with 4 stations and finite constraint T_1 and T_2 jobs. In addition, the stations are paired in 4 pairs (see figure 1.29) with a finite capacity constraint Z_i for the total number of jobs z_i at pair i, by

$$z_i = n_{2i-1} + n_{2i} \leq Z_i$$

(The natural assumption is made that $T_1 \leq Z_1 + Z_2$ and $T_2 \leq Z_3 + Z_4$)As before, cluster balance is violated at cluster 1 if $t_2 = T_2$ and at cluster 2 if $t_1 = T_1$. In addition, station balance is violated at station $2i$ if Z_{i+1} is reached, $i = 1,2,3$. The following modification is therefore is therefore suggested:

PF-modification

- When $t_2 = T_2$ stop stations 1-4
- When $t_1 = T_1$ stop stations 5-8
- When $z_i = n_{2i-1} + n_{2i} = Z_i$:
  stop arrivals and all stations $j \neq 2i-1, 2i$; $i = 1, \ldots, 41-4$

Under this modification the global balance equations become

$$\left\{ \begin{array}{l} \sum_{j=1,\ldots,4} \pi(\boldsymbol{n})\mu_j f_j(n_j) \left[\prod_{i \neq d(j)} 1_{(z_i<Z_i)}\right] \left[1_{(t_2<T_2)}\right] + \\ \sum_{j=5,\ldots,8} \pi(\boldsymbol{n})\mu_j f_j(n_j) \left[\prod_{i \neq d(j)} 1_{(z_i<Z_i)}\right] \left[1_{(t_1<T_1)}\right] + \\ \pi(\boldsymbol{n})\lambda \left[\prod_{i=1}^{4} 1_{(z_i<Z_i)}\right] \left[\prod_{j=1}^{2} 1_{(t_j<T_j)}\right] \end{array} \right\}$$
$$=$$
$$\left\{ \begin{array}{l} \sum_{j=2,\ldots,4} \pi(\boldsymbol{n}+e_{j-2}-e_j)\mu_{j-1} f_{j-1}(n_{j-1}+1) \left[\prod_{i \neq d(j)} 1_{(z_i<Z_i)}\right] \left[1_{(t_2<T_2)}\right] + \\ \sum_{j=5,\ldots,8} \pi(\boldsymbol{n}+e_{j-1}-e_j)\mu_{j-1} f_{j-1}(n_{j-1}+1) \left[\prod_{i \neq d(j)} 1_{(n_i<N_i)}\right] \left[1_{(t_1<T_1)}\right] + \\ \pi(\boldsymbol{n}+e_8)\mu_8 f_8(n_8+1) \left[\prod_{i=1}^{4} 1_{(z_i<Z_i)}\right] \left[\prod_{j=1}^{2} 1_{(t_j<T_j)}\right] \end{array} \right\} \quad (1.106)$$

Here in the right hand side, for $j = 1$ we need to read $n + e_{j-1} + e_j = n - e_1$, $\mu_0 = \lambda$ and $f_0(n_0 + 1) = 1$ and $d(j)$ denotes the pair number that contains station j, $j = 1, \ldots, 8$.

For each $j = 1,\dots,8$ in the first two terms in both hand sides as well as for the third term (as can be seen as for $j = 0$), the indicator functions are identical. Hence, station balance for each j separately is directly verified as before by substituting the product form with $J = 8$ at C the set of admissible states:

$$\begin{aligned} \boldsymbol{S_U} = \{\boldsymbol{n} \mid & t_1 \le T_1 \ , \ t_2 \le T_2 \ , \ t_1 + t_2 \ne T_1 + T_2 \ , \\ & z_i \le Z_i \ , \ i = 1,\dots,4 \ ; \ z_i + z_j \ne Z_i + Z_j \text{ for all } i,j \text{ with } i \ne j\} \end{aligned} \quad (1.107)$$

Clearly, this modification leads to an upper bound $\boldsymbol{B_U}$ for the loss probability $\boldsymbol{B}$. Conversely, a lower bound $\boldsymbol{B_L}$ is obtained by the modification:

PF-modification

- Only reject arrivals when the total number of jobs $t = n_1 + \cdots + n_8 = T_1 + T_2$, while any station can accommodate up to this number of jobs.

In this case again the station balance relations are readily verified with the same product-form as in (1.105) with $\lambda_i = \lambda$ for all i at the set of admissible states::

$$\boldsymbol{S_L} = \{\boldsymbol{n} \mid t \le T_1 + T_2 \ , \ n_i \le T_1 + T_2 \text{ for } i = 1,\dots,4\}$$

Some numerical results are presented in table 1.3.

Table 1.3: Result for the nested blocking structure ($\lambda = 1$)

μ_1	μ_2	μ_3	μ_4	μ_5	μ_6	μ_7	μ_8	$\boldsymbol{Z_1}$	$\boldsymbol{Z_2}$	$\boldsymbol{Z_3}$	$\boldsymbol{Z_4}$	$\boldsymbol{T_1}$	$\boldsymbol{T_2}$	$\boldsymbol{B_L}$	$\boldsymbol{B_U}$	$\boldsymbol{B_{av}}$	$\boldsymbol{B}$
1	1	1	1	1	1	1	1	3	2	4	2	4	5	.471	.724	.598	.572
2	3	4	5	1	2	3	4	3	2	4	2	4	5	.158	.398	.278	.204

Again, this nested assembly line example is also extendable to restricted Jacksonian clusters instead of restricted pairs using the recycle or stop protocol as in section 1.6.

1.7.4 Further illustrative examples

In this section, some more examples will be provided to illustrate the potential of the modification approach. For each of these examples there is no analytic expression known while the modifications guarantee closed product form expressions similar to (1.99). These in turn will lead to easily computable bounds similar to (1.100) and (1.101). Some numerical results will be included to indicate a possible practical usefulness.

1.7.4.1 A Cluster With Parallel Stations

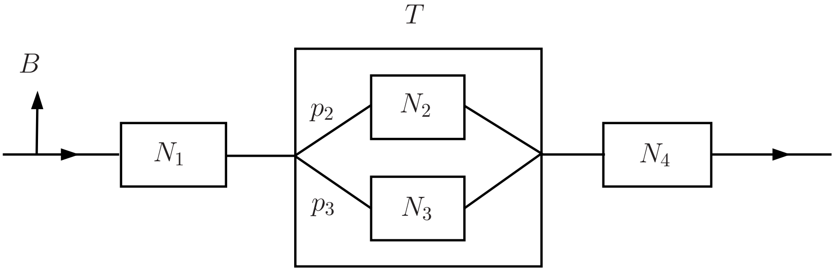


Fig. 1.30: A parallel routing cluster.

This example contains a random routing after completion at station 1 to either one of two stations (with probabilities p_2 and $p_3 = 1 - p_2$) in parallel within one cluster with a capacity constraint $\boldsymbol{T}$ for the total number of jobs at stations 2 and 3, next to capacity constraints N_i at each station i, $i = 1, \ldots, 4$. By regarding the cluster as one aggregated station as in section 2, the following modifications lead to product-form expressions:

PF-modification

- Stop arrivals and all stations either when one of stations ($n_i = N_i$) or the cluster ($n_2 + n_3 = \boldsymbol{T}$) is saturated, or
- Stop arrivals when the total number of jobs is equal to $N_1 + \boldsymbol{T} + N_4 = \boldsymbol{S}$, while each station may contain up to $\boldsymbol{S}$ jobs.

Clearly, the first modification leads to an upper bound $\boldsymbol{B}_U$ and the second to a lower bound B_L for the loss probability $\boldsymbol{B}$ of the original system. Some numerical results are shown by table 1.4.

Table 1.4: Results for the finite cluster with parallel stations ($\lambda = 1$)

$\mu_1 = \ldots = \mu_4$	$\boldsymbol{N}_1$	$\boldsymbol{N}_2$	$\boldsymbol{N}_3$	$\boldsymbol{N}_4$	$\boldsymbol{T}$	$p_2 = p_3$	$\boldsymbol{B}_L$	$\boldsymbol{B}_U$	$\boldsymbol{B}_{av}$	$\boldsymbol{B}$
2	2	2	2	2	3	0.5	.03	.30	.16	.16
10	2	2	2	2	4	0.5	.00	.02	.01	.01
1	5	5	5	5	10	0.5	.10	.30	.18	.20
1	10	5	5	10	10	0.75	.06	.17	.12	.10

1.7.4.2 An Overflow Example

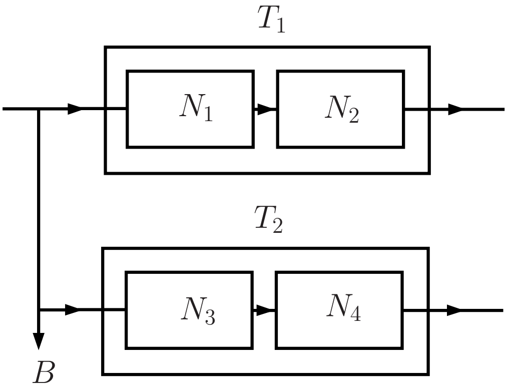


Fig. 1.31: Overflow clusters.

Consider two finite clusters in parallel with arrivals at cluster 1. If a job cannot enter cluster 1 it is rerouted to cluster 2. Each cluster consists of two finite stations in tandem. In addition to the total cluster constraints T_1 and T_2, we also allow capacity constraints N_i for each individual station i, $i = 1,\ldots,4$. We assume that $\mu_1 \le \mu_3$ and $\mu_2 \le \mu_4$.

For this example, the so-called notion of cluster balance is violated when cluster 2 is busy while cluster 1 is not saturated. In that case the outflow at cluster 2 is positive, but the in-rate is 0. The following two modifications are therefore suggested:

PF-modification

- Stop both stations in cluster 2 when cluster 1 is not saturated ($t_1 < \boldsymbol{T}_1$), or
- Assign arriving jobs randomly to either one of the clusters proportional to the free buffer capacity at the two clusters.

By the first modification cluster 2 is slowed down and kept more congested. The arrival loss probability will thus be enlarged which leads to an upper bound $\boldsymbol{B}_U$ for the loss probability $\boldsymbol{B}$ of the original system. With the second modification, the faster overflow cluster is used more frequently than in the original system, which leads to a lower bound $\boldsymbol{B}_L$.

1.7.4.3 A breakdown Model

Reconsider two finite clusters in tandem, which are both subject to breakdowns. In addition to the cluster constraints T_1 and T_2, we assume repair and breakdown rates γ_{10} and γ_{11} for cluster 1, and similarly, γ_{20} and γ_{21} for cluster 2.

Table 1.5: Results for parallel finite clusters with overflow.

λ	μ_1	μ_2	μ_3	μ_4	$\boldsymbol{N}_1$	$\boldsymbol{N}_2$	$\boldsymbol{N}_3$	$\boldsymbol{N}_4$	$\boldsymbol{T}_1$	$\boldsymbol{T}_2$	$\boldsymbol{B}_L$	$\boldsymbol{B}_U$	$\boldsymbol{B}_{av}$	$\boldsymbol{B}$
1	1	1	1	1	1	1	1	1	2	2	.095	.444	.270	.300
2	1	1	4	4	3	3	1	1	6	2	.005	.174	.090	.073
3	1	1	4	4	3	3	2	2	6	4	.023	.126	.075	.063

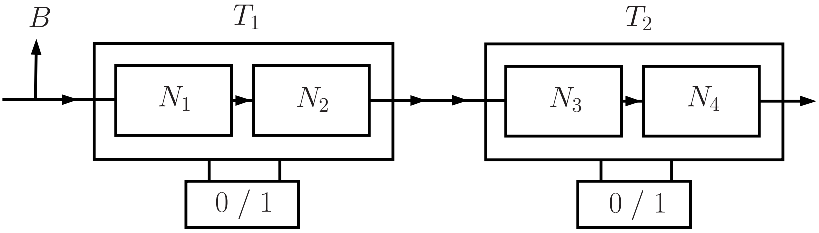


Fig. 1.32: Breakdown clusters.

Clearly, cluster balance is violated when either cluster is down. The following two modifications are therefore suggested:

PF-modification

- Stop both stations in cluster *i* when cluster *j* is down ($j \neq i$), or
- The breakdown rate for both clusters is 0 (breakdowns do not take place).

Again, the first modification leads to an upper bound $\boldsymbol{B}_U$ and the second to a lower bound $\boldsymbol{B}_L$ for the loss probability $\boldsymbol{B}$ of the original system. Some numerical results are shown below.

Table 1.6: Results for finite clusters with breakdowns ($\lambda = 1$)

μ_1	μ_2	μ_3	μ_4	$\boldsymbol{N}_1$	$\boldsymbol{N}_2$	$\boldsymbol{N}_3$	$\boldsymbol{N}_4$	$\boldsymbol{T}_1$	$\boldsymbol{T}_2$	γ_{10}	γ_{11}	γ_{20}	γ_{21}	$\boldsymbol{B}_L$	$\boldsymbol{B}_U$	$\boldsymbol{B}_{av}$	$\boldsymbol{B}$
2	2	2	2	2	2	1	1	4	4	50	1	50	1	.04	.42	.23	.20
2	1	2	1	2	4	2	4	6	6	50	1	50	1	.16	.48	.32	.28

1.7.5 An Optimal Design Application

Reconsider the finite cluster tandem example from section 1.7.1 in which the numbers T_1 and T_2 are still be determined by trading off capacity costs $(T_1 + T_2)^2$ and opportunity losses $1000\boldsymbol{B}$ due to rejections. Based on the lower and upper bounds for the loss probability, lower and upper bound curves for the costs are easily computed. Despite the large discrepancy between the lower and upper bound values, the qualitative curving behavior seems to almost pinpoint the same optimal number (9 or 10). To be more certain one can then simulate. In any case one can be 100% sure that the optimal number is within the region 4-16.

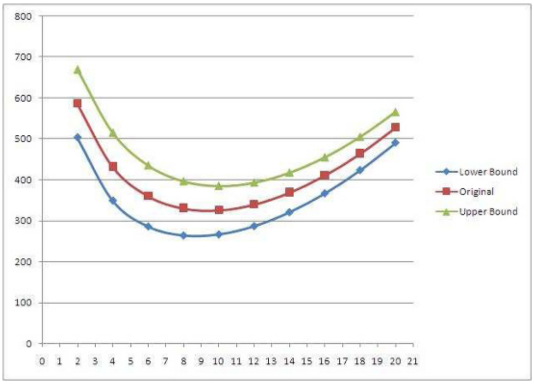


Fig. 1.33: Total capacity optimization.

1.8 A hospital application

1.8.1 Motivation

At an Intensive Care Unit (ICU) within hospitals patients may enter directly for intensive care, such as monitoring and artificial ventilation. Patients may also require an ICU bed for postoperative care after a heavy operation at the Operating Theatre (OT). Unfortunately, due to the limited number of beds, a request for an ICU bed may be rejected.

For patients a rejection may lead to further delay in a critical situation which may even put lives at risk. For the hospital (or public health) a rejection may lead to an idle operating room, which is regarded as a loss of precious capacity. The size of an ICU thus needs to be dimensioned carefully.

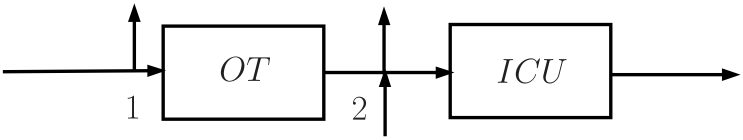


Fig. 1.34: Operating Theatre (OT) - Intensive Care Unit (ICU) Tandem Model

A careful estimation of the ICU rejection probability is thus required. Unfortunately, measurements might not be available or be sufficiently predictive for different number of beds. An analytic approach therefore will be of practical interest.

Literature and objectives. By a number of references the standard $M|G|c|\cdot$ multiserver queue has already been argued as a reasonable approximate for the ICU in isolation (see [55] and references therein). Nevertheless, these results do not contain:

- A formal justification.
- The inclusion of the OT and its interaction with the ICU.
- A secure lower and upper bound for the ICU-rejection probability.

1.8.2 Model formulation

The inflow of the ICU consists of emergency patients (the majority) and elective patients and can be subdivided into various patient groups. However, as we are particularly interested in the effect of the limited ICU capacity and its interaction with the OT, below we only make a cross distinction in patients, that need to visit the ICU after having undergone an operation, and patients that enter the ICU directly without operation. These patients will be referred to as:

- OT (or type 1-) patients.
- Direct (or type 2-) patients.

This distinction is made:

- To capture the interaction between OT and ICU.
- As the average sojourn times at the ICU significantly differ.

Original Model. To study the ICU-rejection probability $\boldsymbol{R}$ for type-1 and type-2 patients (where we refer to remark 1.8.2 below for its equality for both types)and its interaction with the OT a number of assumptions are made. In [55] each of these assumptions has been argued and justified by simulation to be quite reasonable for practical modeling. The corresponding tandem queue system under these assumptions (1)-(8) will be referred to as the *original* OT-ICU model.

(1) Patients that do not require an ICU bed are not included.
(2) A Poisson arrival rate λ_1 of OT-patients (type 1) at the OT.
(3) A Poisson arrival rate λ_2 of Direct patients (type 2) at the ICU.
(4) An exponential service time for the surgery at the OT with rate μ_1.
(5) A (possibly non-exonential) sojourn time at the ICU with mean τ_1 for OT-patients and τ_2 for Direct patients.
(6) The OT has c_1 identical operating rooms with a infinite waiting facility; The ICU has a limited capacity for at most c_2 patients and no waiting facility.
(7) When no ICU bed is available, type 1-patients are rejected upon arrival at the OT and type 2-patients are rejected upon arrival at the ICU.
(8) An ongoing operation is always continued. When no ICU bed is available, the patient is kept in the recovery.

Modified Product Form OT-ICU system. The OT-ICU system of interest has no product form solution. However, in line with the results from section 1.5, more precisely example 1.5.3 and its product form modification in section 1.5.1, the following artificial modification of (8) can be suggested:

(8') When the ICU becomes congested, operations are immediately interrupted and stopped. The operations are resumed as soon as the ICU is no longer congested.

Under this modification, the tandem system will be referred to as the *modified* OT-ICU system. Similarly to the relations (1.64), for this modified OT-ICU system the following result can be proven directly.

Result 1.8.1 *Let $(n_1;m_1,m_2)$ denote that there are n_1 patients at the OT and m_i patients at the ICU of type i $(i=1,2)$. For the modified OT-ICU system, with $m = m_1 + m_2 \leq c_2$,*

$$F_1(n_1) = \begin{cases} [n_1!]^{-1} & \text{for } n_1 \leq c_1 \\ \left[c_1!\, c_1^{(n_1-c_1)}\right]^{-1} & \text{for } n_1 > c_1 \end{cases}$$

and with normalizing constant α, we have:

$$\pi(n_1;m_1,m_2) = \alpha\, F_1(n_1) \left(\frac{\lambda_1}{\mu_1}\right)^{n_1} \prod_{i=1,2} \frac{1}{m_i!} (\lambda_i \tau_i)^{m_i} \tag{1.108}$$

The product form expression (1.108) decomposes as if the OT and ICU can be regarded as independent. As a consequence, with $c = c_2$ it thus directly justifies an $M|G|c|c$-loss approximation for the ICU rejection probability $\boldsymbol{R}$ as by

$$\boldsymbol{B}(c) = \rho^c/c! \left[\sum_{k=0}^{c} \rho^k/k!\right]^{-1} \quad with \;\; \rho = (\lambda_1\tau_1 + \lambda_2\tau_2) \tag{1.109}$$

Intuitively, as the modified OT-ICU tandem system only differs from the original OT-ICU tandem system for a patient in operation when the ICU becomes congested, one may expect that the $M|G|c|c$-loss expression, is a quite reasonable if not accurate approximation for the original OT-ICU system.

Indeed, as shown in table XXX below, this $M|G|c|c$-loss approximation for the ICU rejection probability R of the OT-ICU as described in section 1.8.1 and practically argued by simulation in [55], seems to approximate quite well with c in the order of the case study as in section 1.8.3.

TABLE

Remark 1.8.2 *The ICU-rejection probability for type-1 patients is equal to that for the type-2 patients. This can be argued directly by the PASTA (Poisson Arrivals See Time Averages) property. Alternatively, it can also be concluded from the product form result 1.8.1 by (1.108).*

1.8.3 Bounds and application

Nevertheless, there is no guarantee at all for this approximation to be accurate. More importantly, one might intuitively expect that it provides a lower bound $\boldsymbol{B}(c) \leq \boldsymbol{R}$, as the modification seems to keep operations more conservative. In practice, in contrast one would rather have a secure upper bound, such as for dimensioning the size of an ICU with a secure sufficiently small rejection probability (e.g. less than 5%). Here in addition it is also noted that the Poisson arrival assumption for type 1 jobs is somewhat unrealistic, as operations are partly scheduled, and thus overestimates the 'practical' value $\boldsymbol{R}$. An upper rather than lower bound for $\boldsymbol{R}$ would thus be of interest. Based upon the product form modification again as in section 1.8.2, but with $c-1$ rather than c beds and a Markov reward proof technique as outlined in section a chapter later on the following result is therefore proven in [55].

Result 1.8.3 (Bounds for the ICU-rejection probability)

$$\boldsymbol{B}(c) \leq \ \boldsymbol{R} \ \leq \boldsymbol{B}(c-1) \tag{1.110}$$

Application: Case study. Data were collected for a case study in a Dutch hospital over a one year period. The percentages of type 1 and type 2-patients were 39% and 61%.

The average sojourn time spend in the ICU over all patients was 5.2 days, for roughly 4 days for type 1-patients and 6 days for type 2-patients. Other case characteristics were:

- OT capacity (number of operating rooms): 8.

- ICU capacity (number of beds): 12.
- ICU occupancy: 85%.

The case study situation is within a range of realistic figures as recently reported by the Dutch ministry of health. It reported that roughly 10% of ICU requests are strictly rejected, 3% admitted by a predischarge and 4% placed differently. Furthermore an occupancy of 75% is mentioned as norm.

Simulation results for the case study consistently support the lower and upper bound. Particularly, for smaller rejection probabilities, say in the order of 5 -10% as for larger hospitals with a high occupancy level, the bounds appear to be quite accurate (in absolute sense). The results seem useful, at least, for practical purposes such as to guarantee a sufficiently small rejection percentage by the upper bound.

For the case study, an occupancy of 85% and 12 beds were used. The results lead to a lower bound of .127 and an upper bound of .172 (the simulation result was .128). As a direct application of the secure $M|G|c$-$1|c$-1 upper bound computation the required number of ICU beds could be computed as:

- 16 beds for $\boldsymbol{R} \leq 5\%$.
- 19 beds for $\boldsymbol{R} \leq 1\%$.

1.9 Evaluation

1.9.1 Literature

Product forms for queueing networks have become most familiar ever since the pioneering work of Jackson ([28], [29]) for so-called Jackson networks. In these networks a job can randomly route form one station to another. Some most notable other early references here are [30], [36], [37] and [35]. Particularly, for its clear practical motivation, special attention has been given to assembly line structures, also referred to as Gordon-Newell networks ([18], [19]). In fact, an early first product form result for a production line system was already reported in [36], [37].

In none of these references, though, the product forms were 'explicitly' related to (a notion of) balance for each station separately (In the elegant paper [35] an explicit decomposition is made for the outside as a station).

A verification by each station separately as if it were in isolation became more emphasized in the seventies in [3], [10], [11], [33], [34], [35], [44], [47], [48] (with corresponding balance notions as local, detailed balance and job-local balance).

With (access) blocking due to finite capacity constraints product form results seemed more restricted. In fact the historical paper by Jackson does already contains a capacity constraint on the total number of jobs in the network. But with finite capacity constraints N_i for the total number of jobs at individual stations i, product form results were concluded only provided the routing is reversible [33], [45]. As mentioned in both these references, in this case also arbitrary truncation of the state space can be incorporated while retaining a product form. In fact, the results in section 1.2 on coordinate convex blocking and the corresponding references as [11], [31], [39] fall within this category as the routing for entering and leaving a single service station is reversible by definition. [33] and [46] also include a reversible result along this line as a job has a designated service path through the network for which it is either entirely accepted (that is the whole path) or rejected. An intermediate blocking or stagnation is not allowed.

The product form papers by [3], [10], [11], [44] which focus on multiple-class extensions of Jackson networks as well as different queueing disciplines, which are well-known in the computer communication literature (as also discussed in section 1.4.2, **A**) do not allow for any blocking by finite capacity constraints at all. Extensions of such networks with blocking but strictly with a reversible routing can be found in [33], [53], [67].

However, even for a simple two stage tandem model, as in example 1.5.2, the routing is necessarily not reversible. As a consequence, a simple finite constraint as by an intermediate buffer as in example 1.5.3, violates a product form.

In view of the practical importance of tandem (or assembly line) structures, extensive attention has therefore been paid to approximations for such structures, as in [14], [16], [17], [22], [49] and most recently [62], [63].

The approach taken in this chapter is different in that it aims to investigate the existence of product forms by more general blocking functions for two reasons:

- As of interest by itself to investigate to which extent product forms can still be concluded also for non-reversible blocking or service sharing
- As motivated by the product form modification example under example 1.5.3. This modification was shown to provide simple and practical bounds [56], [59]. Further exploration of a product form bounding approach is thus of practical interest.

The notion of adjoint reversibility and its product form characterization as by (1.65)-(1.69) in section 1.5.2 and the necessary and sufficient blocking condition (1.83) in section 1.5.3 for the tandem example, have essentially been developed in [23]. An extension of this characterization to arbitrary multi-class Jackson type networks with a job-configuration (state) dependent routing and servicing can be found in [25]. This characterization is restricted in [50] to single-class networks with a state dependence on the configuration vector $\boldsymbol{n}$ for the total number of jobs at each station. The presentation in section 1.5.2, more precisely the characterizations (1.69) and (1.70), as well as (1.71)-(1.76) in remarks 1.5.9 and 1.5.10 directly rely upon this reference. This also applies to the blocking examples in section 1.5.4 while the (mixed) service examples 1.5.14, 1.5.18 and 1.5.20 can be regarded as 'new, though included in [51] and [61]. Other product form examples of interest which fit in this framework but which merely focus on service sharing over stations, such as for internet modelling, can be found in recent work by [4] and [61].

A total network constraint N for the total number of jobs as in section 1.6.2 was already incorporated in the original famous paper by Jackson [28]. The Jacksonian product form results in sections 1.6.2 and 1.6.3 can be concluded from [50] and [52] but are presented here in a self-contained compact form. The bounding results for networks with restricted clusters in section 1.7 are adopted from [57], but are also presented in a compact self-contained form. In this reference an analytic representation and 'cluster balance' relation are given to consider a restricted cluster as one aggregate station. (For a more general setting of product form results and the possibility of some form of decomposition or aggregation of stations, as extension of Norton's theorem, the interested reader is also referred to [8] and [5], [6]). The special hospital application in section 1.8 is obtained from [54], as based on the research paper [55].

Finally, for more detailed discussions and reference lists on product form results in more abstract lists on product form results in more abstract settings, most notably as in [12], [46], [64] the reader is referred to the chapters by Miyazawa, by Daduna, and by Boucherie and Huisman in this book.

1.9.2 Review Part B

In Part B of this chapter just single-class tandem-type structures were considered, in its most simple form with just two successive stations. Such simple structures might already be regarded as generic for a variety of application fields, as in manufacturing for production lines, as in communications for internet and packet switch networks, or as in service environments like a hospital.

In contrast with Part A, however, the focus in Part B has been the phenomenon of blocking by finite capacity limitations at stations or by service sharing between stations, due to either of which successive stations essentially become interdependent. Even in the simple two-station case this dependence may render the system unsolvable.

Nevertheless, by assuming general state dependent blocking and service functions, product form results could still be concluded. These results were essentially based upon the three types of:

(i) Requiring station balance equations.

(ii) A translation of these equations into an adjoint chain.

(iii) A product form characterization by means of reversibility for this adjoint chain (called adjoint reversibility).

This characterization led to concrete product form examples also with blocking such as by a ’conservative’ or ’recycle’ blocking ’protocol’. Though these product form examples might if not will generally be unrealistic, the underlying station balance insights and explicit product form expressions might still be practically useful to provide simple performance bounds.

This statement seemed supported by extensions and numerical results for more complex tandem structures (with Jacksonian clusters) as well as a realistic hospital case. Further extension and application of these product form insights and a bounding approach is thus suggested.

1.9.3 Some remaining questions

Despite the general perception that product form results are exhaustively covered in the literature, to the opinion of the author, a variety of intriguing questions, which are also of practical interest, remain open for research. Three of them and far from exhaustible are:

1. To what extent can the detailed product form results as in **A** for a single station simply be embedded in a more global (blocking) network structure as in **B**. Only without blocking (or dependence phenomena) or under special conditions as a reversible routing some results for mixed networks with different types of stations have been reported .
2. Somewhat related to 1 and in line with decomposition and aggregation results (as in [5], [8]), can we also recognize and come up with closed (product) form expressions for just a subpart of a network, despite the fact the network in totality is unsolvable.
3. The bounding approach as used in section 1.7 is strongly supported at a physical and intuitive level as by a strict blocking or service interruption to let an inflow or outflow become 0. The recognition of product form modifications will be a less transparent if non-zero modifications have to be found, such as for resolving unproportional processor sharing, as of present-day interest for internet applications modeling (fairness).

Please feel free to join these product form questions.

Acknowledgements

The author likes to acknowledge Richard Boucherie for his persistent motivation to write this chapter. He is highly indebted to Nikky Kortbeek for his critical reading and his careful assistance in editing this chapter from which the presentation has greatly benefitted.

The support by the Dutch Organisation for Scientific Research (NWO) by which this research could be realized has been highly appreciated.

References

1. Adan, I.J.B.F, Wessels, J. (1993). Product Forms as a Solution Base for Queueing Systems. *Operations Research Proceedings 1992*, Springer-Verlag, Berlin, 316-323.
2. Barbour, A. (1976). Networks of queues and the method of stages. *Advances in Applied Probability*, **8**, 584-591.
3. Baskett, F., Chandy, M., Muntz, R. and Palacios, J. (1975). Open, closed and mixed networks of queues with different classes of customers. *JACM*, **22**.
4. Bonald, T. and Proutiere, A. (2002). A Queueing Analysis of Max-min Fairness, Proportional Fairness and Balanced Fairness. *Queueing Systems*, **53(1-2)**, 193-209.
5. Boucherie, R.J. (1992). *Product-Form in Queueing Networks*. PhD dissertation, Vrije Universiteit, Amsterdam. Tinbergen Institute Research Series 28, Thesis Publishers Amsterdam.
6. Boucherie, R.J., Huisman, T. (1999). A State-dependent Generalisation of Quasi-reversibility and Biased Local Balance in Queueing Networks, Report AE 3/99, Institue of Actuarial Sciences & Econometrics, University of Amsterdam.
7. Boucherie, R.J., and Van Dijk, N.M. (1991). Product forms for queueing networks with statedependent multiple job transitions. *Adv. Appl. Prob.*, **23**, 152-187.
8. Boucherie, R.J., and Van Dijk, N.M. (1993). A generalization of Norton's Theorem. *Queueing Systems*, **13**, 251-289.
9. Brockmeyer, E., Halstrom H.L., Jensen, A. (1948). *The Life and Works of A.K. Erlang*. Copenhagen: The Copenhagen Telephone Company.
10. Chandy, K.M., and Martin, A.J. (1983). A Characterization of Product-Form Queueing Networks. *JACM*, **30**, 2, 286-299.
11. Chandy, K.M., Howard, J.H. and Towsley, D.F. (1977). Product form and local balance in queueing networks. *JACM*, **24**, 250-263.
12. Chao, X., Miyazawa, M., Pinedo, M. (1999). *Queueing Networks: Customers, Signals and Product form Solutions*. Wiley, Chichester.
13. Conway, A.E. (1989). Product form and insensitivity in circuit-switched networks with failing links. *Performance Evaluation*, **9**, 209-215.
14. Dallery, Y. and Frein, Y. (1989). A decomposition method for the approximate analysis of closed queueing networks with blocking. *Queueing Networks with Blocking (Eds Perros, H.G. and Altiok, T.)*, North-Holland, 193-215.
15. Foshini, G.J. and Gopinath, B. (1983). Sharing memory optimally. *IEEE Trans. Comm.*, **31**, 352-359.
16. Gershwin, S.B. (1978). An efficient decomposition method for the approximate evaluation of tendem queues with finite storage space and blocking. *Operations Research*, March-April, 291-305.
17. Gershwin, S.B. (1989). An efficient decomposition algorithm for unreliable tandem queueing systems with finite buffers. *Queueing Networks with Blocking (Eds Perros, H.G. and Altiok, T.)*, North-Holland, Amsterdam.
18. Gordon, W.J. and Newell, G.F. (1967a). Cyclic queueing systems with restricted length queues. *Operations Research*, **15**, 266-277.
19. Gordon, W.J. and Newell, G.F. (1967b). Closed queueing systems with exponential servers. *Operations Research*, **15**, 254-265.
20. Henderson, W. and Taylor, P. (1988). Alternative routing network and interruptions. *ITC*, **12**, Torino, 5.IB.2.1.
21. Heyman, D.P. (1980). Comments on a queueing inequality. *Management Science*, **26**, 956-959.

22. Hiller, F.S. and Boling, R.W. (1976). Finite queues in series with exponential or Erlang service times - a numerical approach. *Operations Research*, **15**, 286-303.
23. Hordijk, A., and Van Dijk, N.M. (1981). Networks of queues with blocking. *Performance (Ed. Kylstra, K.J.)*, **81**, North-Holland, 51-65.
24. Hordijk, A., and Van Dijk, N.M. (1982). Stationary probabilities for networks of queues. *Applied Probability-Computer Science: The Interface (eds. Disney , L., and Ott, T.J.)*, Brikhuser, Vol. II, 423-451.
25. Hordijk, A., and Van Dijk, N.M. (1983). Networks of queues, Part I: Job-local-balance and the adjoint process Part II: General routing and service characteristics. *Lecture in Notes in Control and Information Sciences*, **60**, Springer-Verlag, 158-205.
26. Hordijk, A., and Van Dijk, N.M. (1983). Adjoint processes, job-local-balance and insensitivity for stochastic networks. *Bull 44th Session Int. Stat.*, **50**, 776-788.
27. Inose, H., Kato, M. and Saito, T. (1967). No-hole-in-the-multiple alternate routing system. *Electro Comm. Japan*, **50**, No. 11, 11.
28. Jackson, J.R. (1957). Networks of waitinglines. *Operations Research*, **5**, 518-521.
29. Jackson, J.R. (1963). Jobshop-like queueing systems. *Management Sciences*, **10**, 131-142.
30. Jackson, R.P.P. (1954). Queueing systems with phase-type service. *Operat. Res. Quart.* , **5**, 109-120.
31. Kaufman, J. (1981). Blocking in a shared resource environment. *IEEE Trans. Comm.*, **29**, 1471-1481.
32. Kelly, F.P. (1975). Networks of queues with customers of different types. *J. Appl. Probability*, **12**, 542-554.
33. Kelly, F.P. (1979). *Reversibility and Stochastic Networks*, Wiley.
34. Kelly, F.P. (1976). Networks of queues. *Adv. Appl. Probability*, **8**, 416-432.
35. Kingman (1969). Markov population process. *J. Appl. Probability*, **6**, 1-18.
36. Koeningsberg, E. (1958). Cyclic queues. *Operational Research: Quarterly*, **9**, 22-35.
37. Koeningsberg, E. (1959). Production Lines and Internal Storage. A Review. *Management Science*, **9**, 410-433.
38. Kroese, D.P., Scheinhardt, W.R.W., and Taylor, P.G. (2004). Spectral Properties of Tandem Jackson Network, seen as a Quasi-Birth-and-Death Process. *Annals of Applied Probability*, **14**, no. 4.
39. Lam, S.S. (1977). Queueing networks with population size constraints. *IBM J. Res. Development*, 370-378.
40. Latouche, G. and Neuts, M.F. (1980). Efficient algorithmic solutions to exponential tandem queues with blocking. *SIAM J. Alg. Disc. Meth.*, **1**, 93-105.
41. Lipper, E.H. and Sengupta, B. (1986). Assembly-like queues with finite capacity: bounds, asymptotics and approximations. *Queuing Systems*, **1**, 67-83.
42. Miyazawa, M., Taylor, P.G. (1997). Geometric product-form distribution for a queueing network with nonstandard batch arrivals and batch transfers. *Advances in Applied Probability*, **29**, 523-544.
43. Molina, E.C. (1927). Application of theory of probability to telephone trunking problems. *Bell System Tech. Journal*, **6**, 461.
44. Noetzel, A.S. (1979). A generalized queueing discipline for product form network solutions. *JACM*, **26**, 4, 779-793.
45. Pittel, B. (1979). Closed Exponential Networks of Queues with Saturation. The Jackson-type Stationary Distribution and its Asymptotic Analysis. *Math. Of Operations Res.*, **4**, 367-378.
46. Serfozo, R. (1989). Markovian network process: congestion dependent routing and processing. *Queueing Systems*, **5**, 5-36.
47. Schassberger, R. (1978). The Insensitivity of Stationary Probabilities in Networks of Queues. *Adv. Appl. Prob.*, **10**, 906-912.
48. Schassberger, R. (1979). A definition of Discrete Product Form Distribution in Networks of Queues. *Zeitschrift fur Operations Research*, **23**, 189-195.
49. Shanthikumar, J.G. and Jafari, M. (1987). Bounding the performance of tandem queues with finite spaces, *Technical report*, School of business Administration, University of California, Berkeley.

50. Van Dijk, N.M. (1993). *Queueing Networks and Product Forms: A Systems Approach*, Wiley, Chichester.
51. Van Dijk, N.M. (2005). On Product Form Tandem Structures. *Mathematical Methods of Operations Research*, **62(3)**, 429-436.
52. Van Dijk, N.M. (1988). On Jackson's Product Form with "Jump-over" Blocking. *Oper. Res. Letters*, **7**, 233-235.
53. Van Dijk, N.M. and Akyildiz, I.F. (1990). Networks with mixed processor sharing parallel queues and common pools. *Performance*, **90**, North-Holland, Amsterdam, 35-49.
54. Van Dijk, N.M., Kortbeek, N. (2008). On dimensioning Intensive Care Units. *Operations research, Proceedings 2007, Selected Papers of the Annual Internation Conference of the German Operations Research Society (GOR), Saarbrücken, September 5-7*, 291-296.
55. Van Dijk, N.M., Kortbeek, N. (2009). Erlang Loss Bounds for OT-ICU systems. *To appear: Queueing systems*, **67**.
56. Van Dijk, N.M., Lamond, B.F. (1988). Bounds for the Call Congestion of Finite Single-server Exponential Tandem Queues. *Operations Research*, **36**, 470-477.
57. Van Dijk, N.M., Van der Sluis, H.J (2004). Simple Product-Form Bounds for Queueing Networks with Finite Clusters. *Annals of Operations Research*, **113**, 175-195.
58. Van Dijk, N.M., Van der Sluis, H.J. (2008). Call Packing Bounds for Overflow Queue. *Performance Evaluation*, To appear.
59. Van Dijk, N.M., Van der Wal, J. (1989). Simple Bounds and Monotonocity Results for Finite Multi-server Exponential Tandem Queues. *Queueing Systems*, **4**, 1-16.
60. Van Dijk, N.M., Tsoucas, P., Walrand, J. (1988). Simple Bounds and Monotonocity of the Call Congestion of Infinite Multiserver Delay Systems. *Probability in the Engineering and Informational Sciences*, **2**, 129-138.
61. Van der Weij, W., Van Dijk, N.M., Van der Mei, R.D. (2008). Research Report.
62. Van Vuuren, M., Adan, I.J.B.F. (2007). Approximate analysis of general priority queues. *Proceedings of Analysis of Manufacturing Systems*, 139-145.
63. Van Vuuren, M., Adan, I.J.B.F. (2006). Performance analysis of assembly systems. *Proceedings of the Markov Anniversary Meeting* 89-100.
64. Van Vuuren, M., Adan, I.J.B.F., Resing-Sassen, S.A.E. (2005). Performance analysis of multi-server tandem queues with finite buffers. *OR Spektrum*, **27**, 315-338.
65. Walrand, J. (1988). *Introduction to Queueing Systems*. Wiley, New York.
66. Whittle, P. (1986). *Systems in Stochastic Equilibrium*, Wiley, New York.
67. Yao, D.D. and Buzacott, J.A. (1987). Modelling a class of flexible manufacturing systems with reversible routing. *Operations Research*, **35**, 87-93.

Chapter 2
Order Independent Queues

A.E. Krzesinski

Abstract We present a class of queues which are quasi-reversible and therefore preserve product form distribution when connected in multinode networks. The essential feature leading to the quasi-reversibility of these queues is the fact that the total departure rate in any queue state is independent of the order of the customers in the queue. We call such queues Order Independent (OI) queues. A distinguishing feature of the OI class is that, among others, it includes the FCFS, processor sharing, infinite server and MSCCC queues but not the LCFS queue. We next examine OI queues where arrivals to the queue are lost when the number of customers in the queue equals an upper bound. We prove that such queues satisfy partial balance and we obtain the stationary distribution for the OI loss queue by normalising the stationary probabilities of the corresponding OI queue without losses. OI loss queues can be used to model systems with simultaneous resource possession with the option of queueing blocked customers. The OI loss queue thus extends previous loss models where customers are rejected when processing resources are not available. The OI loss class is next extended to include networks of queues which can be used to model systems with complex loss mechanisms. We finally present several applications of OI loss queues and OI loss networks.

2.1 Introduction

Much attention has been given to product form stationary distributions for queueing networks and associated Markov processes – see [17, 25, 30, 32] for a list of references. Most of the known processes which have a product form distribution are reversible or quasi-reversible. Quasi-reversibility was first presented by Muntz [24]

A.E. Krzesinski
Department of Mathematical Sciences, University of Stellenbosch, 7600 Stellenbosch, South Africa
e-mail: aek1@cs.sun.ac.za

R.J. Boucherie and N.M. van Dijk (eds.), *Queueing Networks: A Fundamental Approach*,
International Series in Operations Research & Management Science 154,
DOI 10.1007/978-1-4419-6472-4_2, © Springer Science+Business Media, LLC 2011

and later developed into a general framework [17, 29]. In the product form Jackson networks [16], each node in isolation is a reversible Markov process. In the more general BCMP product form networks [3], nodes may not be reversible, but they are quasi-reversible.

However, product form is not restricted to systems of interconnected quasi-reversible nodes. For example, Pollett [27, 28] provided a framework for interconnecting a collection of reversible Markov processes in such a way that the resulting process has a product form invariant measure with respect to which the process is reversible, although individual nodes need not be quasi-reversible. Kelly [17] developed a general framework for interconnecting quasi-reversible processes and introduced a class of quasi-reversible symmetric queues that can be connected in a network with a product form distribution. Walrand [29, 30, 31] provided probabilistic arguments demonstrating how quasi-reversibility implies product form distribution. One of the most general frameworks for interconnecting quasi-reversible nodes [14] does not require the node to be customer preserving for each type of customer.

Other quasi-reversible queues are often variations of reversible queues [17, 30] and can be described as reversible queues by choosing an appropriate state space for the queue. There are other (unclassified) quasi-reversible queues [10, 17, 18, 30, 32] and processes such as the quasi-reversible Brownian process considered in [13], Ott's model analysed in [14] and quasi-reversible clustering processes [32].

This paper presents a class of quasi-reversible queues which in general are neither symmetric nor reversible. This class includes a large part of the class of symmetric queues, but not the whole class. In particular, the well known FCFS, Infinite-Server and Processor-Sharing queues are in the considered class which also contains the Multiserver Station with Concurrent Classes of Customers (MSCCC) [7, 11, 12] and the MultiServer centre with Hierarchical Concurrency Constraints (MSHCC) [21]. The quasi-reversibility for this class is not a result of the symmetry of the reversed process as is the case for symmetric queues. It arises from a symmetry property concerning the order of the customers in the queue, namely that the total departure rate in any queue state is independent of the order of the customers in the queue. We call such queues Order Independent (OI) queues.

In section 2.2 we define the OI queue using the notation presented in [17]. This allows us to simplify the comparison of the OI queue with other well known quasi-reversible queues and to emphasise the distinguishing features of the OI queue. The OI property is obtained by analysing the properties of so-called service rate functions which satisfy some special conditions. In section 2.2.3 we prove that if the instantaneous service rate of a multiclass queue is described by a set of such service rate functions, then the queue is quasi-reversible (in the appropriate state space). We derive a closed form expression for the stationary distribution of the queue.

In section 2.2.4 we show that an appropriate choice of relative service rate functions reduces the OI queue to the well known BCMP, MSCCC and MSHCC queues. Another OI queue which is a generalisation of example 3 from [18] is presented. Section 2.3 derives a computationally efficient recursive equation for the stationary distribution. Section 2.4 examines OI queues with losses. We prove that although OI loss queues are not quasi-reversible, their stationary distributions can be obtained by normalising the stationary probabilities of the corresponding OI queue without losses. Finally, section 2.5 presents several applications of OI loss queues and OI loss networks.

Much of the work presented in this Chapter was done jointly with Sergei Berezner whom I wish to thank for his many valuable insights and discussions on the subject.

2.2 The OI Queue

Consider a queue serving customers of type c where $c \in \mathcal{C}$ and $\mathcal{C}$ is a finite set. Customers of type c arrive individually at the instants of a Poisson stream with rate λ_c. The customers, whether waiting or in service, form a queue in the order of their arrival. Arriving customers join the back of the queue and the front of the queue is identified with position 1. Each customer of type c presents a demand for service time which is exponentially distributed with mean $1/\mu_c$. All the random variables involved in the description of the queue are independent.

Let $\mathbf{C} = (c_n, \ldots, c_1)$ denote the state of a queue of length n where c_i denotes the type of the customer in position i, $i = 1, \ldots, n$. Let 0 denote the empty queue and $\mathcal{S} = \{0\} \cup \bigcup_{n=1}^{\infty} \mathcal{C}^n$ denote the state space of the queue where $\mathcal{C}^n$ is the $n-$fold product space of $\mathcal{C}$.

Let the total service effort in state $(c_n, \ldots, c_1)$ be supplied at the rate $\phi(c_n, \ldots, c_1)$. A portion $\gamma_i(c_n, \ldots, c_1)$ of the total service effort is directed at the customer in queue position i, $i = 1, \ldots, n$. Upon entering service a customer is served without interruption to completion. When the customer in queue position i completes service, the customer departs and the gap in the queue is closed by the obvious shift: the customers in positions $i+1, i+2, \ldots, n$ move to positions $i, i+1, \ldots, n-1$ respectively. We do not require that $\sum_{i=1}^{n} \gamma_i(c_n, \ldots, c_1) = 1$ so that a part of the service facility might be wasted. This is a distinguishing feature of an OI queue and this feature allows complex queueing disciplines such as MSCCC and MSHCC to be described as OI queues.

Note that the OI queue is completely described by the vector of types of customers. For some specific cases it may be possible to introduce auxiliary variables that describe additional characteristics of the queue and for such queues the concept of OI could be further extended. However, this will result in a cumbersome notation without extending the range of the models in the OI class.

We therefore restrict ourselves to queues that are completely described by a vector of types of customers. We also assume that the type membership of each customer does not change while it passes through the queue. This is a severe restriction since it prevents us from considering service consisting of a series of exponential stages and, as a result, general service distributions are not possible for OI queues. This restriction can be dropped for some specific models, for example, for PS or IS queues. But in general, service in stages does not lead to an OI queue and typical OI queues such as M/M/K, MSCCC and MSHCC do not permit service in stages.

2.2.1 The Definition of an OI Queue

We next present a general description of the OI queue and show how the OI property can be obtained by imposing certain conditions on the functions ϕ and γ.

First, we require that the (relative) proportion of the service effort supplied to the customer in queue position i depends only on the composition of the queue up to and including position i. This implies that when a server becomes free the queue is searched from the head to the tail for a customer which can be admitted into service.

Second, we require that in any state $(c_n,\ldots,c_1) \in \mathcal{S}$, the rate at which departures (service completions) occur is independent of the order of the customers in the queue and is thus the same for any state $(c_{\sigma(n)},\ldots,c_{\sigma(1)})$, where σ denotes any permutation of $(1,\ldots,n)$.

Last, we assume that in any state (except for the empty queue) there is a positive rate of service completion. This condition is required in order to ensure the irreducibility of the Markov chain. This restriction is usually satisfied in systems with exponential service that are fully described by a vector of customer types who do not change their type membership.

A formal description of the three conditions presented above can be given as follows. Consider the queue in state $(c_n,\ldots,c_1)$. The departure rate of the customer in queue position i is given by $\phi(c_n,\ldots,c_1)\mu_{c_i}\gamma_i(c_n,\ldots,c_1)$ and the total departure rate is given by the sum of these quantities over all the positions in the queue.

The queue is said to be an OI queue if, for all $(c_n,\ldots,c_1) \in \mathcal{S}$ and all $i = 1,\ldots,n$ the rates of service completion can be written as

$$\phi(c_n,\ldots,c_1)\mu_{c_i}\gamma_i(c_n,\ldots,c_1) = \mu(n)s_i(c_n,\ldots,c_1)$$

such that

(i) $s_i(c_n,\ldots,c_1) = s_i(c_i,\ldots,c_1)$ for any $1 \le i \le n$,

(ii) $k(c_n,\ldots,c_1) = \sum_{i=1}^{n} s_i(c_n,\ldots,c_1)$ is independent of permutations of $(c_n,\ldots,c_1)$, and

(iii) $\mu(n) > 0$ for $n > 0$ and $s_1(c) > 0$ for any $c \in \mathcal{C}$.

The function $s_i(\mathbf{C})$ regulates the rate at which service is given to the customer in position i in the queue relative to the other customers in the queue and the function $\mu(n)$ allows the service rate to depend upon the total number of customers in the queue.

2.2.2 The Implications of the OI Conditions

Condition (i) requires that the relative service rate of any customer in the queue depends only upon its own customer type and the customer type of each customer in front of it in the queue. This implies that it is only necessary to scan the queue from the front to determine the relative service rate given to any particular customer.

Condition (ii) is the distinguishing condition. It requires that the total departure rate due to customer service completions is independent of the order of the customers in the queue. Although this condition is restrictive, section 2.2.4 shows that condition (ii) is satisfied by several well known queues.

Condition (iii) is a necessary and sufficient condition for the Markov process describing the queue to be irreducible. This condition ensures that the empty state can always be reached from any other state due to departures. Because arrivals allow any state to be reached from the empty state, this is sufficient to ensure irreducibility. Note that because $S_i(\mathbf{C}) \geq 0$, a necessary and sufficient condition for $k(\mathbf{C}) > 0$ for any $\mathbf{C} \in \mathcal{S}$ is to require that $s_1(c) > 0$ for any $c \in \mathcal{C}$.

Theorem 2.1. *Conditions (i) and (ii) imply that the relative service rate given to a customer in the queue is independent of the order of the customers ahead of it in the queue.*

Proof. Let $(c_n,\dots,c_1)$ be a queue state in $\mathcal{S}$ and let $(\sigma(1),\dots,\sigma(n-1))$ denote a permutation of $(1,\dots,n-1)$. Thus $(c_n,c_{\sigma(n-1)},\dots,c_{\sigma(1)})$ is the queue state obtained when the first $n-1$ customers in the queue are rearranged according to the permutation $(\sigma(1),\dots,\sigma(n-1))$. By definition

$$k(c_n,\dots,c_1) = \sum_{i=1}^{n} s_i(c_n,\dots,c_1) = s_n(c_n,\dots,c_1) + k(c_{n-1},\dots,c_1)$$

and likewise

$$k(c_n,c_{\sigma(n-1)},\dots,c_{\sigma(1)}) = s_n(c_n,c_{\sigma(n-1)},\dots,c_{\sigma(1)}) + k(c_{\sigma(n-1)},\dots,c_{\sigma(1)}). \quad (2.1)$$

But condition (ii) requires that $k(c_n,c_{\sigma(n-1)},\dots,c_{\sigma(1)}) = k(c_n,\dots,c_1)$ and $k(c_{\sigma(n-1)},\dots,c_{\sigma(1)}) = k(c_{n-1},\dots,c_1)$. Substituting these two expressions into (2.1) yields $s_n(c_n,\dots,c_1) = s_n(c_n,c_{\sigma(n-1)},\dots,c_{\sigma(1)})$ which completes the proof. □

Note that the $s_i(c_i,\dots,c_1)$ may be dependent on the permutations of $(c_i,\dots,c_1)$, but not on permutations of $(c_{i-1},\dots,c_1)$. However $k(c_n,\dots,c_1)$ is not dependent on permutations of $(c_i,\dots,c_1)$. For example, consider an OI queue with two customer types where only one customer from each type may be in service. Then, for example, $s_3(2,1,1) = 1$ whereas $s_3(1,2,1) = s_3(1,1,2) = 0$ (the last two instances of $s_3(c_3,c_2,c_1)$ illustrate theorem 2.1) and $k(2,1,1) = k(1,2,1) = k(1,1,2) = 2$.

The following section demonstrates that the restrictions (i)–(iii) on $s_i(\mathbf{C})$ are sufficient to ensure that the OI queue is quasi-reversible at equilibrium and we find the stationary distribution (when it exists).

2.2.3 The Stationary Distribution

The OI queue can be modeled by a continuous–time, homogeneous Markov process $\mathbf{C}(t)$, $t \in \mathbf{R}^+$, $\mathbf{C}(t) \in \mathcal{S}$, where $\mathbf{C}(t)$ denotes the queue state at time t. The Markov process is irreducible, since transitions due to arrivals allow any state to be reached from the empty state and condition (iii) ensures that the empty state can be reached from any state by the departure transitions.

For any $c \in \mathcal{C}$ and any $(c_n,\dots,c_1) \in \mathcal{S}$ the transition rate due to type c arrivals is λ_c and the transition rate due to the departure of a type c customer in queue position $i+1$ where $1 \le i \le n$ is $\mu(n+1)s_i(c_n,\dots,c_{i+1},c,c_i,\dots,c_1)$. The equilibrium equations for the queue are given by

$$\pi(0)\sum_{c\in\mathcal{C}}\lambda_c = \mu(1)k(c)\sum_{c\in\mathcal{C}}\pi(c) \tag{2.2}$$

and

$$\begin{aligned}&\pi(c_n,\dots,c_1)\left(\lambda+\mu(n)k(c_n,\dots,c_1)\right)\\ &= \sum_{c\in\mathcal{C}}\sum_{i=0}^{n}\mu(n+1)\pi(c_n,\dots,c_{i+1},c,c_i,\dots,c_1)s_{i+1}(c_n,\dots,c_{i+1},c,c_i,\dots,c_1)\\ &+\quad \lambda_{c_n}\pi(c_{n-1},\dots,c_1)\end{aligned} \tag{2.3}$$

where $\lambda = \sum_{c\in\mathcal{C}}\lambda_c$.

The arrival flow to the system is a collection of independent Poisson flows each with rate λ_c with future arrivals being independent of the present state of the queue. If we can find a collection of positive numbers $\pi(c_n,\dots,c_1)$ summing to unity and satisfying the equilibrium Eqs. (2.2) and (2.3) such that for all $(c_n,\dots,c_1) \in \mathcal{S}$ and all $c \in \mathcal{C}$

$$\sum_{i=0}^{n}\frac{\pi(c_n,\dots,c_{i+1},c,c_i,\dots,c_1)}{\pi(c_n,\dots,c_1)}\mu(n+1)s_{i+1}(c,c_i,\dots,c_1) = \beta_c \tag{2.4}$$

then the queue is quasi-reversible and this collection of numbers forms a stationary distribution of the queue [18]. Since customers in an OI queue preserve their type membership, $\beta_c = \lambda_c$. Thus (2.4) can be rewritten as

$$\sum_{i=0}^{n}\frac{\pi(c_n,\dots c_{i+1},c,c_i,\dots,c_1)}{\pi(c_n,\dots,c_1)}\mu(n+1)s_{i+1}(c,c_i,\dots,c_1) = \lambda_c. \tag{2.5}$$

Note that condition (i) was applied in (2.4) to replace $s_{i+1}(c_n,\dots,c_{i+1},c,c_i\dots,c_1)$ by $s_{i+1}(c,c_i,\dots,c_1)$. Substituting (2.5) into (2.3) yields

$$\mu(n)k(c_n,\dots,c_1)\pi(c_n,\dots,c_1) = \lambda_{c_n}\pi(c_{n-1},\dots,c_1) \tag{2.6}$$

for all $(c_n,\dots,c_1) \in \mathcal{S}$. From (2.6) we immediately obtain a proposed form of the stationary distribution. The result is stated and proved in the following theorem.

Theorem 2.2. *If the service rate functions $s_i(\cdot)$ conform to the conditions (i)–(iii) then for any $(c_n,\dots,c_1) \in \mathcal{S}$ a collection of numbers*

$$\pi(c_n,\ldots,c_1) = \pi(0)\prod_{i=1}^{n}\frac{\lambda_{c_i}}{\mu(i)k(c_i,\ldots,c_1)} \tag{2.7}$$

where $\pi(0)$ is an arbitrary positive real number, is a solution to the equilibrium equations. The stationary distribution of the Markov chain exists if and only if

$$G = \sum_{(c_n,\ldots,c_1)\in\mathcal{S}}\prod_{i=1}^{n}\frac{\lambda_{c_i}}{\mu(i)k(c_i,\ldots,c_1)} < \infty$$

in which case the stationary distribution is given by (2.7) with $\pi(0) = 1/G$ and the queue is quasi-reversible.

Proof. Equation (2.7) is clearly a solution to (2.6). We need to prove that (2.7) is also a solution to (2.5) and thus the solution to the equilibrium equations. We prove by induction on n that the equality (2.5) holds for all $(c_n,\ldots,c_1)\in\mathcal{S}$ and $c\in\mathcal{C}$.

Consider first the empty queue $(n=0)$. Applying (2.7) to the right hand side of (2.5), and noting that $k(c)\equiv s_1(c)$, yields

$$\frac{\mu(1)\pi(c)s_1(c)}{\pi(0)} = \frac{\mu(1)\pi(0)s_1(c)}{\lambda_c\pi(0)\mu(1)k(c)} = \lambda_c$$

so that the base of the induction is proved.

Next assume that (2.7) satisfies (2.5) up to the $n-1$ value of the summation index and consider (2.5) for the value n of the summation index. The left hand side of (2.5) (lhs (2.5)) can be written as

$$\begin{aligned}\text{lhs }(2.5) &= \sum_{i=0}^{n}\frac{\pi(c_n,\ldots,c_{i+1},c,c_i,\ldots,c_1)}{\pi(c_n,\ldots,c_1)}\mu(n+1)s_{i+1}(c,c_i,\ldots,c_1)\\ &= \sum_{i=0}^{n-1}\frac{\pi(c_n,\ldots,c_{i+1},c,c_i,\ldots,c_1)}{\pi(c_n,\ldots,c_1)}\mu(n+1)s_{i+1}(c,c_i,\ldots,c_1)\\ &\quad+\frac{\pi(c,c_n,\ldots,c_1)}{\pi(c_n,\ldots,c_1)}\mu(n+1)s_{n+1}(c,c_n,\ldots,c_1).\end{aligned}$$

Application of (2.7) yields

$$\begin{aligned}\text{lhs }(2.5) = &\frac{\mu(n)k(c_n,\ldots,c_1)}{\mu(n+1)}\\ &\times\sum_{i=0}^{n-1}\frac{\pi(c_{n-1},\ldots,c_{i+1},c,c_i,\ldots,c_1)}{\pi(c_{n-1},\ldots,c_1)k(c_n,\ldots,c_{i+1},c,c_i,\ldots,c_1)}\mu(n+1)s_{i+1}(c,c_i,\ldots,c_1)\\ &+\frac{\lambda_c}{k(c,c_n,\ldots,c_1)}s_{n+1}(c,c_n,\ldots,c_1).\end{aligned}$$

Condition (ii) requires that $k(\mathbf{C})$ is independent of permutations of $\mathbf{C}$, hence we obtain $k(c_n,\ldots,c_{i+1},c,c_i,\ldots,c_1) = k(c,c_n,\ldots,c_1)$. Condition (iii) ensures that $k(c,c_n,\ldots,c_1)\mu(n+1) > 0$. Thus

$$\text{lhs }(2.5) = \frac{k(c_n,\ldots,c_1)}{k(cc_n\ldots c_1)}\sum_{i=0}^{n-1}\frac{\pi(c_{n-1},\ldots,c_{i+1},c,c_i,\ldots,c_1)}{\pi(c_{n-1},\ldots,c_1)}\mu(n)s_{i+1}(c,c_i,\ldots,c_1)$$
$$+\frac{\lambda_c s_{n+1}(c,c_n,\ldots,c_1)}{k(c,c_n,\ldots.c_1)}.$$

But by the induction assumption the sum on the right hand side of the above equation is equal to λ_c , which yields

$$\text{lhs }(2.5) = \lambda_c\frac{k(c_n,\ldots,c_1)+s_{n+1}(c,c_n,\ldots,c_1)}{k(c,c_n,\ldots,c_1)} = \lambda_c$$

which completes the induction. □

2.2.4 Models Covered by the OI Class

In this section we show that the OI class includes several of the BCMP queues and part of Kelly's class of symmetric queues. A distinguishing feature of the OI queues is that they include the MSCCC and the MSHCC queues.

2.2.4.1 BCMP Models in the OI Class

The M/M/K queue is an OI queue. Let $\mu(n) = 1$ and

$$s_i(c_n,\ldots,c_1) = \begin{cases} \mu & 1 \le i \le K \\ 0 & i > K \end{cases}$$

for $K \in \mathbf{Z}^+$. These functions $\mu(n)$ and $s_i(\cdot)$ conform to the conditions (i)–(iii) and this OI queue is equivalent to an M/M/K queue. All customer types must have the same average service rate else $k(\mathbf{C})$ would not be independent of permutations of $\mathbf{C}$. We will later show that the M/M/K queue is a special case of the MSCCC queue. This implies that the MSCCC queue could replace the M/M/K queue as a basic construction element (building block) for product form networks.

We next consider the Infinite Server (IS) queue. Let $\mu(n) = 1$ and $s_i(c_n,\ldots,c_1) = \mu_{c_i}$. These functions $\mu(n)$ and $s_i(\cdot)$ conform to the conditions (i)–(iii) and this OI queue is equivalent to an IS queue.

The choice of relative service rates for the Processor Sharing (PS) queue is obvious. Let $\mu(n) = 1/n$ and $s_i(c_n,\ldots,c_1) = \mu_{c_i}$. These functions $\mu(n)$ and $s_i(\cdot)$ conform to the conditions (i)–(iii) and the corresponding OI queue is equivalent to a PS queue. As was mentioned in section 2.2, it is possible to extend the OI framework to describe general service distributions for PS and IS queues. However, this extension complicates the proofs and does not lead to any fundamentally new models since in this case the OI conditions (i)–(iii) become very restrictive. In any event, typical OI queues such as M/M/K, MSCCC and MSHCC do not admit general service distributions.

Note that the PS and IS queues require all customers to be in service and that in these cases the service effort $k_c(\mathbf{C})$ directed at customers of type c given by

$$k_c(c_n,\ldots,c_1) = \sum_{i=1}^{n} s_i(c_n,\ldots,c_1)\, 1(c_i = c)$$

is also independent of permutations of $\mathbf{C}$. In general, it is not necessary that all customers be in service at an OI queue with type dependent service rates. Service disciplines may exist for OI queues with type dependent service rates with a bound on the number of customers in service – all that is necessary for such a queue to be OI is that conditions (i)–(iii) hold.

Because of condition (i) the Last Come First Served (LCFS) queue cannot be modeled by an OI queue. The failure of the OI class to describe the LCFS queue is another argument proving the different nature of the OI class.

If the functions ϕ and γ defined in section 2.2 depend only on the total number of customers in the queue and if we allow the customers to join the queue in different positions then, under some strong symmetric assumptions (see [17] pages 72–73), the well known symmetric queues are obtained.

2.2.4.2 The MSCCC and MSHCC Queues

A distinguishing feature of the OI class is that it includes the MSCCC and MSHCC queues. In fact, the OI queues were found while investigating the MSCCC queue.

The MSCCC queue consists of K parallel identical exponential servers. The customers belong to a set of customer types. Customers of type c arrive individually at the instants of a Poisson stream with rate λ_c and present demands for service time which are exponentially distributed with mean $1/\mu$. The customers are queued for service in the order of their arrival. When a server becomes free, the queue is searched from the front looking for the first customer to admit into service subject to the following constraints: at most K customers can be in service and at most 1 customer of each type c can be in service. The queue can thus be described as FCFS subject to concurrency constraints.

The MSCCC queue was first investigated while simulating shared memory multiprocessors [22]. The simulations conjectured, but could not prove, that the MSCCC queue had a product form solution. An analytic expression for the stationary distribution, which was first obtained by exploring the MSCCC state space using a symbolic mathematics computer package, was presented in [7]. The concurrency constraint was later extended [11] so that at most K customers can be in service and at most $B_c \geq 1$ customer of each type c can be in service. It was later shown [21] that a suitable choice of state descriptor yields a direct calculation of the stationary distribution. However, none of these analyses explained how the MSCCC queue is related to the product form queues.

The OI property presents a simple explanation of the MSCCC queue which is shown to be a quasi-reversible generalisation of the FCFS queue. Let $\mu(n) = 1$ and

$$s_i(c_n,\dots,c_1) = \mu 1_i(c_n,\dots,c_1)$$

where the indicator function

$$1_i(c_n,\dots,c_1) = \begin{cases} 1 & \text{if the customer in position } i \text{ is in service in } (c_n,\dots,c_1) \\ 0 & \text{otherwise} \end{cases}$$

Then

$$k(c_n,\dots,c_1) = \mu(K \wedge \sum_{c\in\mathcal{C}} (M_c \wedge B_c))$$

where M_c is the number of type c customers in $\mathbf{C} = (c_n,\dots,c_1)$ and $a \wedge b$ is the smaller of the two integers a and b. As required, $k(\mathbf{C})$ is independent of permutations of $\mathbf{C}$ and the functions $\mu(n)$ and $s_i(\cdot)$ conform to the conditions (i)–(iii). Furthermore, if not all K servers are busy, $k_c(\mathbf{C}) = M_c \wedge B_c$ is independent of permutations of $\mathbf{C}$.

Theorem 2.2 presents the stationary distribution of the MSCCC queue: the analysis is simpler than the theorems currently available [7, 11] and provides further insight into the behaviour of the MSCCC queue.

The constraints $(B_c)_{c\in\mathcal{C}}$ can be further divided into an hierarchical structure of concurrency constraints. Thus the set of customer types $\mathcal{C}$ is partitioned as $\{\mathcal{C}_r; r \in \mathcal{R}\}$ where $\mathcal{R}$ is a countable set and (with an abuse of notation) maximally $B_r > 0$ customers whose types are in $\mathcal{C}_r$ are allowed to be in service simultaneously where $r \in \mathcal{R}$. Next, for each $r \in \mathcal{R}$ the set $\mathcal{C}_r$ is partitioned as $\{\mathcal{C}_{rs}; s \in \mathcal{S}_r\}$ where $\mathcal{S}_r$ is a countable set and maximally $B_{rs} > 0$ customers whose types are in $\mathcal{C}_{rs}$ are allowed to be in service simultaneously where $s \in \mathcal{S}_r$. The $\mathcal{C}_{rs}$ can be further partitioned, with corresponding restrictions placed on the number of customers simultaneously in service, but we shall not go beyond the $\mathcal{C}_{rs}$ as this is a straightforward generalization. The queue discipline can thus be described as FCFS subject to concurrency constraints and the queue is correspondingly named the MultiServer centre with Hierarchical Concurrency Constraints (MSHCC) [21]. Then

$$k(c_n,\dots,c_1) = \mu(K \wedge \sum_{r\in\mathcal{R}} (m_r \wedge B_r))$$

and

$$m_r = \sum_{s\in\mathcal{R}_r} (m_{rs} \wedge B_{rs})$$

where m_{rs} is the number of $\mathcal{C}_{rs}$ customers in $\mathbf{C} = (c_n,\dots,c_1)$. The MSHCC queue conforms to the conditions (i)–(iii) which immediately provides us with the stationary distribution.

Another example of an OI queue is provided by example 3 from [18]. This queue consists of K servers with service rates $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_K$. The customers, which belong to a single customer type, are processed according to the following rule.

If j servers are busy then these j busy servers are the servers $1,\dots,j$. If $j<K$ then an arriving customer will be served by server $j+1$, else the arrival joins the back of the queue. Let M denote the total number of customers in the queue. If a customer completes service at any one of the busy servers $i \in 1,\dots,j$ where $1 \leq j \leq M \wedge K$ then the customer at the end of the queue in position M will be removed from processor j (only if $M<K$ in which case $M=j$) and go into service at server i (only if $i<K$). The queue is OI and the queue can be applied to model dynamic load balancing among asymmetric multiprocessors.

Summarising, the OI conditions (i)–(iii) do not leave much room for a wide range of applications. On the other hand, the fact that several well known quasi-reversible queues are OI queues speaks in favour of the OI discipline. Another valuable feature of OI queues which we prove in the next section is that they permit a standard normalising technique for calculating the stationary distributions of OI systems.

2.3 Numerical Techniques for the OI Queue

This section presents several techniques which are used in the numerical analysis of OI queues.

2.3.1 Aggregating the State Space

Equation (2.7) is too detailed to be of practical use when computing the performance measures of the OI queue. In order to reduce the complexity of these equations, we use the fact that $k(c_n,\dots,c_1)$ is independent of the order of the customers in $(c_n,\dots,c_1)$.

Define a mapping $A : \mathcal{S} \mapsto \boldsymbol{Z}^C$ such that for any $\mathbf{C} \in \mathcal{S}$

$$A(\mathbf{C}) = \mathbf{M}(\mathbf{C}) = (M_c(\mathbf{C}))_{c \in \mathcal{C}}$$

where M_c denotes the number of type c customers in $\mathbf{C}$. In the remainder of this section, where no confusion can arise we shall write $\mathbf{M} = \mathbf{M}(\mathbf{C})$. Let $A^{-1}(\mathbf{M})$ denote the set of elements in $\mathcal{S}$ which maps onto $\mathbf{M}$ under A. The mapping A allows us to define an aggregated state space

$$\mathcal{M} = \{\mathbf{M} = A(\mathbf{C}),\, \mathbf{C} \in \mathcal{S}\}$$

where (with an abuse of notation)

$$k(\mathbf{M}) = k(\mathbf{C})$$

for any $\mathbf{C} \in A^{-1}(\mathbf{M})$ and (with an abuse of notation)

$$\pi(\mathbf{M}) = \sum_{A(\mathbf{C})=\mathbf{M}} \pi(\mathbf{C})$$

for any $\mathbf{M} \in \mathcal{M}$. Because condition (ii) ensures that $k(\mathbf{C})$ is identical for all $\mathbf{C} \in A^{-1}(\mathbf{M})$, $k(\mathbf{M})$ is well defined and unique.

Let $\mathbf{1}_c$ denote a unit vector in the c-direction. From (2.7) it follows that

$$\mu(|\mathbf{M}|)k(\mathbf{M})\pi(\mathbf{M}) = \sum_{c\in\mathcal{C}} \lambda_c \pi(\mathbf{M} - \mathbf{1}_c) \tag{2.8}$$

with the convention that $\pi(\mathbf{M}) = 0$ if $\mathbf{M} \notin \mathcal{M}$.

The recursive (2.8) can be further simplified if the service effort $k_c(\mathbf{M})$ directed at customers of type c for any $\mathbf{M} \in \mathcal{M}$ can be determined. Although condition (ii) requires that $k(\mathbf{C})$ is independent of permutations of $\mathbf{C}$, this does not necessarily imply that $k_c(\mathbf{C})$ is independent of permutations of $\mathbf{C}$. However, in the domain where $k_c(\mathbf{C})$ is identical for all $\mathbf{C} \in A^{-1}(\mathbf{M})$, we can define

$$k_c(\mathbf{M}) = k_c(\mathbf{C})$$

for any $\mathbf{C} \in A^{-1}(\mathbf{M})$ in which case (2.8) reduces to

$$\mu(|\mathbf{M}|)k_c(\mathbf{M})\pi(\mathbf{M}) = \lambda_c \pi(\mathbf{M} - \mathbf{1}_c)$$

which provides a recursion for the efficient computation of the aggregated probabilities and the performance measures of the queue.

2.3.2 The Performance Measures: the MSCCC Queue

Unlike the BCMP queues, efficient recursions for the MSCCC queue apply only in a limited portion of the state space where $k(\mathbf{M}) < B$.

2.3.2.1 Invariant Measures over Special Sets

Let $\mathbf{M} = (M_1, \ldots, M_C)$ denote the state of the queue where M_c is the number of type c customers present in the queue. Let B_c denote the maximum number of type c customers in service and B the maximum number of customers in service in total. Define the sets

$$\begin{aligned}
\mathcal{M}(b) &= \{\, \mathbf{M} \in \mathcal{M} \mid k(\mathbf{M}) = b \,\} \\
\mathcal{M}(b,c) &= \{\, \mathbf{M} \in \mathcal{M}(b) \mid M_x = 0 \text{ if } x > c \,\} \\
\mathcal{M}(b,c,i) &= \begin{cases} \{\, \mathbf{M} \in \mathcal{M}(b,c) \mid M_c = i \,\} \text{ if } i < B_c \\ \{\, \mathbf{M} \in \mathcal{M}(b,c) \mid M_c \geq i \,\} \text{ if } i = B_c. \end{cases}
\end{aligned}$$

Define the functions

$$\begin{aligned} P(b) &= \Pr(\,b \text{ servers busy}\,) \\ P(b,c) &= \Pr(\,b \text{ servers busy and } M_x = 0 \text{ if } x > c\,) \\ P(b,c,i) &= \begin{cases} \Pr(\,b \text{ servers busy, } M_x = 0 \text{ if } x > c \text{ and } M_c = i\,) \text{ if } i < B_c \\ \Pr(\,b \text{ servers busy, } M_x = 0 \text{ if } x > c \text{ and } M_c \geq i\,) \text{ if } i = B_c. \end{cases} \end{aligned}$$

The first step in the calculation of $P(b)$ is to compute the $P(b,c,i)$. The function $k_c(\mathbf{M})$ defined in section 2.3.1 cannot be determined over the complete state space, but in the limited portion of the state space where $k(\mathbf{M}) < B$ where not all servers are busy, we know that $k_c(\mathbf{M}) = M_c \wedge B_c$ and $k(\mathbf{M} - \mathbf{1}_c) = k(\mathbf{M}) - 1$. Define $\rho_c = \lambda_c/\mu$. For any $c \in \mathcal{C}$ and $0 < b < B$ three cases arise

(1) $0 < i < b \wedge B_c$

$$\begin{aligned} P(b,c,i) &= \sum_{\mathbf{M} \in \mathcal{M}(b,c,i)} \pi(\mathbf{M}) \\ &= \sum_{\mathbf{M} \in \mathcal{M}(b,c,i)} \frac{\rho_c}{k_c(\mathbf{M})} \pi(\mathbf{M} - \mathbf{1}_c) \\ &= \frac{\rho_c}{i} \sum_{\mathbf{M} \in \mathcal{M}(b-1,c,i-1)} \pi(\mathbf{M}) \\ &= \frac{\rho_c}{i} P(b-1,c,i-1). \end{aligned} \tag{2.9}$$

(2) $i = B_c$ and $B_c \leq b \leq B$

$$\begin{aligned} P(b,c,B_c) &= \sum_{\mathbf{M} \in \mathcal{M}(b,c,B_c)} \pi(\mathbf{M}) \\ &= \sum_{\substack{\mathbf{M} \in \mathcal{M}(b,c) \\ M_c = B_c}} \pi(\mathbf{M}) + \sum_{\substack{\mathbf{M} \in \mathcal{M}(b,c) \\ M_c > B_c}} \pi(\mathbf{M}) \\ &= \sum_{\substack{\mathbf{M} \in \mathcal{M}(b,c) \\ M_c = B_c}} \frac{\rho_c}{B_c} \pi(\mathbf{M} - \mathbf{1}_c) + \sum_{\substack{\mathbf{M} \in \mathcal{M}(b,c) \\ M_c > B_c}} \frac{\rho_c}{B_c} \pi(\mathbf{M} - \mathbf{1}_c) \\ &= \frac{\rho_c}{B_c} \sum_{\mathbf{M} \in \mathcal{M}(b-1,c,B_c-1)} \pi(\mathbf{M}) + \frac{\rho_c}{B_c} \sum_{\mathbf{M} \in \mathcal{M}(b,c,B_c)} \pi(\mathbf{M}) \\ &= \frac{\rho_c}{B_c} P(b-1,c,B_c-1) + \frac{\rho_c}{B_c} P(b,c,B_c) \\ &= \frac{\rho_c}{B_c - \rho_c} P(b-1,c,B_c-1). \end{aligned} \tag{2.10}$$

(3) $i = 0$

$$\begin{aligned} P(b,c,0) &= \Pr\{\, b \text{ servers busy}, M_x = 0 \text{ if } x > c \text{ and } M_c = 0\,\} \\ &= \Pr\{\, b \text{ servers busy}, M_x = 0 \text{ if } x > c-1\,\} \\ &= P(b,c-1). \end{aligned} \tag{2.11}$$

The starting values of the recursion presented in (2.9), (2.10) and (2.11) are obtained from

$$P(0,c,0) = \begin{cases} 1 \text{ if } c = 0 \\ 0 \text{ otherwise} \end{cases} \tag{2.12}$$

for all $0 \le c \le C$. The next step is to compute the $P(b,c)$

$$P(b,c) = \sum_{i=0}^{b \wedge B_c} P(b,c,i) = P(b,c-1) + \sum_{i=1}^{b \wedge B_c} P(b,c,i) \tag{2.13}$$

where

$$\begin{aligned} P(b,0) &= \Pr\{\, b \text{ servers busy and } M_x = 0 \text{ if } x > 0\,\} \\ &= \Pr\{\, b \text{ servers busy and } \mathbf{M} = \mathbf{0}\,\} \\ &= \begin{cases} 1 \text{ if } b = 0 \\ 0 \text{ otherwise.} \end{cases} \end{aligned}$$

The last step is to compute the $P(b)$

$$\begin{aligned} P(b) &= \Pr\{\, b \text{ servers busy}\,\} \\ &= \Pr\{\, b \text{ servers busy and } M_x = 0 \text{ if } x > C\} \\ &= P(b,C) \end{aligned}$$

which are multiples of $\pi(\mathbf{0})$. To calculate $\pi(\mathbf{0})$, note that when b customers are in service, the departure rate is $b\mu$. Thus $\rho = \sum_{b=0}^{B} bP(b) < B$ where $\rho = \sum_{c=1}^{C} \rho_c$ and $\sum_{b=0}^{B} P(b) = 1$ so that

$$B - \rho = \sum_{b=0}^{B-1} (B-b)P(b). \tag{2.14}$$

Let $\overline{P}(b)$ represent the values obtained by starting the algorithm with an arbitrary value for $P(0,0,0)$ in (2.12). Then

$$\pi(\mathbf{0}) = \frac{B-\rho}{\sum_{b=0}^{B-1} (B-b)\overline{P}(b)} \tag{2.15}$$

where $\overline{P}(0) = 1$. Thus for $0 \le b < B$

$$P(b) = \pi(\mathbf{0})\overline{P}(b) \tag{2.16}$$

and

$$P(B) = 1 - \sum_{b=0}^{B-1} P(b). \tag{2.17}$$

Equations (2.15) and (2.17) do not require the value $P(B)$ which is not available from the recursion presented in (2.9), (2.10) and (2.11).

Application of (2.9), (2.10) and (2.11) yields another recursion for $P(b,c)$ namely

$$\begin{aligned} P(b,c) &= \sum_{i=0}^{B_c} P(b,c,i) \\ &= \sum_{i=0}^{B_c-1} \frac{\rho_c^i}{i!} P(b-i,c-1) + \frac{\rho_c^{B_c}}{B_c!} \frac{B_c}{B_c-\rho_c} P(b-B_c,c-1) \end{aligned} \tag{2.18}$$

where $0 < c \leq C$ and $0 < b < B$ with the convention that $P(x,y) = 0$ if $x < 0$ or $y < 0$. The recursion (2.18) has the same computational complexity namely $O(CB^2)$ as (2.9), (2.10) and (2.11) but with O(B) as opposed to O(B^2) storage requirements. Note that (2.18) does not replace the recursions (2.9), (2.10) and (2.11) which are required in order to compute the expected number of type C customers in the system.

2.3.2.2 The Expected Queue Length

Let $L(b,C)$ denote the expected queue length of type C customers in $\mathcal{M}(b,C)$

$$\begin{aligned}
\frac{B_C}{\rho_C} L(b,C) &= \frac{B_C}{\rho_C} \sum_{\mathbf{M}\in\mathcal{M}(b,C)} M_C \pi(\mathbf{M}) \\
&= \frac{B_C}{\rho_C} \sum_{i=1}^{\infty} \sum_{\substack{\mathbf{M}\in\mathcal{M}(b,C)\\ M_C=i}} i\,\pi(\mathbf{M}) \\
&= B_C \sum_{i=1}^{B_C} \sum_{\substack{\mathbf{M}\in\mathcal{M}(b,C)\\ M_C=i}} \pi(\mathbf{M}-\mathbf{1}_C) + \sum_{i=B_C+1}^{\infty} \sum_{\substack{\mathbf{M}\in\mathcal{M}(b,C)\\ M_C=i}} i\,\pi(\mathbf{M}-\mathbf{1}_C) \\
&= B_C \sum_{i=1}^{B_C} \sum_{\substack{\mathbf{M}\in\mathcal{M}(b,C)\\ M_C=i}} \pi(\mathbf{M}-\mathbf{1}_C) + \sum_{i=B_C+1}^{\infty} \sum_{\substack{\mathbf{M}\in\mathcal{M}(b,C)\\ M_C=i}} \pi(\mathbf{M}-\mathbf{1}_C) \\
&\quad + \sum_{i=B_C+1}^{\infty} \sum_{\substack{\mathbf{M}\in\mathcal{M}(b,C)\\ M_C=i}} (i-1)\,\pi(\mathbf{M}-\mathbf{1}_C) \\
&= B_C \sum_{i=0}^{B_C-1} \sum_{\substack{\mathbf{M}\in\mathcal{M}(b-1,C)\\ M_C=i}} \pi(\mathbf{M}) + \sum_{i=B_C}^{\infty} \sum_{\substack{\mathbf{M}\in\mathcal{M}(b,C)\\ M_C=i}} \pi(\mathbf{M}) \\
&\quad + \sum_{i=B_C}^{\infty} \sum_{\substack{\mathbf{M}\in\mathcal{M}(b,C)\\ M_C=i}} i\,\pi(\mathbf{M}) \\
&= B_C \sum_{i=0}^{B_C-1} P(b-1,C,i) + P(b,C,B_C) \\
&\quad + \sum_{i=0}^{\infty} \sum_{\substack{\mathbf{M}\in\mathcal{M}(b,C)\\ M_C=i}} i\,\pi(\mathbf{M}) - \sum_{i=0}^{B_C-1} \sum_{\substack{\mathbf{M}\in\mathcal{M}(b,C)\\ M_C=i}} i\,\pi(\mathbf{M}) \\
&= B_C \sum_{i=0}^{B_C-1} P(b-1,C,i) + P(b,C,B_C) + L(b,C) - \sum_{i=0}^{B_C-1} iP(b,C,i)
\end{aligned}$$

which yields

$$\frac{(B_C-\rho_C)}{\rho_C} L(b,C) = P(b,C,B_C) + \sum_{i=0}^{B_C-1} \left(B_C P(b-1,C,i) - iP(b,C,i)\right). \tag{2.19}$$

The expected number $L(C)$ of type C customers in the queue is

$$L(C) = \sum_{b=0}^{B} L(b,C). \tag{2.20}$$

Equation (2.20) cannot be used directly to determine $L(C)$ because we do not have a value for $L(B,C)$. However, consider

$$\begin{aligned}
\sum_{b=0}^{B} bL(b,C) &= \sum_{b=0}^{B} b \sum_{\mathbf{M}\in\mathcal{M}(b)} M_C\pi(\mathbf{M}) \\
&= \sum_{b=0}^{B} \sum_{\mathbf{M}\in\mathcal{M}(b)} M_C k(\mathbf{M})\pi(\mathbf{M}) \\
&= \sum_{b=0}^{B} \sum_{\mathbf{M}\in\mathcal{M}(b)} M_C \sum_{c=1}^{C} \rho_c \pi(\mathbf{M}-\mathbf{1}_c) \\
&= \sum_{c=1}^{C} \rho_c \sum_{\mathbf{M}\in\mathcal{M}} M_C \pi(\mathbf{M}-\mathbf{1}_c) \\
&= \sum_{c=1}^{C-1} \rho_c \sum_{\mathbf{M}\in\mathcal{M}} M_C \pi(\mathbf{M}-\mathbf{1}_c) + \rho_C \sum_{\mathbf{M}\in\mathcal{M}} M_C \pi(\mathbf{M}-\mathbf{1}_C) \\
&= \sum_{c=1}^{C-1} \rho_c \sum_{\mathbf{M}\in\mathcal{M}} M_C \pi(\mathbf{M}-\mathbf{1}_c) \\
&\quad + \rho_C \sum_{\mathbf{M}\in\mathcal{M}} (M_C - 1)\pi(\mathbf{M}-\mathbf{1}_C) + \rho_C \sum_{\mathbf{M}\in\mathcal{M}} \pi(\mathbf{M}-\mathbf{1}_C) \\
&= \sum_{c=1}^{C-1} \rho_c L(C) + \rho_C L(C) + \rho_C \\
&= \rho L(C) + \rho_C
\end{aligned}$$

so that

$$\rho L(C) = \sum_{b=0}^{B} bL(b,C) - \rho_C. \tag{2.21}$$

Combining (2.20) and (2.21) yields

$$(B-\rho)L(C) = \sum_{b=1}^{B-1} (B-b)L(b,C) + \rho_C. \tag{2.22}$$

Little's law is applied to obtain the expected time $W_C = L(C)/\lambda_C$ that a type C customer spends at the queue. Equations (2.9) through (2.22) can be incorporated into a Mean Value Analysis algorithm [7, 8, 11, 12] to compute the performance measures for a mixed multiclass network of BCMP and MSCCC queues.

Equation (2.22) yields the type C queue length. The performance measures for any type $c \in \mathcal{C}$ are obtained by reordering the set $\mathcal{C}$. An algorithm to compute the type C performance measures of the MSCCC queue is presented in the Appendix.

2.4 The OI Loss Queue

The class of OI queues can be extended to include queues with complex loss mechanisms. Such queues can be used to model blocking systems with simultaneous resource possession – see [26] for list of references on the application of such models.

Consider a queue serving customers which belong to C customer types. Let $\mathcal{C} = \{1,2,\ldots,C\}$ denote the set of customer types. Customers of type c arrive individually to a queue of length n at the instants of a Poisson stream with rate $\lambda_c r(n)$, where the multiplier $r(n) > 0$ does not depend on c. Each customer of type c presents a demand for service which is exponentially distributed with mean $1/\mu$.

The customers, whether waiting or in service, form a queue in the order of their arrival. A Markov chain with vector states $\mathbf{C} = (c_n,\ldots,c_1)$ is used to describe the queue, where $c_i \in \mathcal{C}$, $i = 1,\ldots,n$ identifies the type of the customer in queue position i and n is the number of customers in the queue. The 0 state of the Markov chain describes the empty queue. Let $\mathcal{C}^n$ denote the $n-$fold product space of $\mathcal{C}$. Then $\mathcal{S} = 0 \cup \bigcup_{n=1}^{\infty} \mathcal{C}^n$ is the set on which the Markov chain is defined.

An arriving customer is either accepted to the queue, in which case it joins the back of the queue, or is rejected (lost) if certain limits on the numbers of customers of each type in the queue are exceeded. The rejection rule for arriving customers is defined as follows. Define a set $\widetilde{\mathcal{S}} \subset \mathcal{S}$ which satisfies the following properties

(a) if $(c_n,\ldots,c_1) \in \widetilde{\mathcal{S}}$ then $(c_{\sigma(n)},\ldots,c_{\sigma(1)}) \in \widetilde{\mathcal{S}}$ for any permutation $\sigma(1),\ldots,\sigma(n)$ of $1,\ldots,n$ and
(b) if $(c_n,\ldots,c_1) \in \widetilde{\mathcal{S}}$ then $(c_{n-1},\ldots,c_1) \in \widetilde{\mathcal{S}}$.

A type c customer arriving to a queue in the state $\mathbf{C}$ will be accepted if $(c,\mathbf{C}) \in \widetilde{\mathcal{S}}$ and rejected otherwise.

Condition (a) implies that acceptance does not depend on the order of the customers in the queue but only on the number of customers of each type present in the queue. To understand condition (b) let $A(\mathbf{C}) = \mathbf{n} = (n_c)_{c\in\mathcal{C}}$ where n_c denotes the number of type c customers in $\mathbf{C}$ and let $\widetilde{\mathcal{N}} = A(\widetilde{\mathcal{S}})$. Conditions (a) and (b) imply that the set $\widetilde{\mathcal{N}}$ is coordinately convex [26] so that $\mathbf{n} \in \widetilde{\mathcal{N}}$ implies $\mathbf{n} - \mathbf{1}_c \in \widetilde{\mathcal{N}}$ where $\mathbf{1}_c$ is a unit vector in the c^{th} direction.

The queue described above is said to be an OI loss queue on the set $\widetilde{\mathcal{S}}$ if for all $\mathbf{C} \in \widetilde{\mathcal{S}}$ and all $i = 1,\ldots,n$ the departure rate of the customer in queue position i can be written as $\mu(n)s_i(c_n,\ldots,c_1)$ where

(i) $s_i(c_n,\ldots,c_1) = s_i(c_i,\ldots,c_1)$ for any $1 \le i \le n$,
(ii) $k(c_n,\ldots,c_1) = \sum_{i=1}^{n} s_i(c_n,\ldots,c_1)$ is independent of permutations of $(c_n,\ldots,c_1)$ and
(iii) $\mu(n) > 0$ for $n > 0$ and $s_1(c) > 0$ for any $c \in \mathcal{C}$.

2.4.1 The Stationary Distribution

The OI loss queue can be modeled by a continuous-time, homogeneous Markov chain $X(t), t \in \mathbf{R}^+, X(t) \in \widetilde{\mathcal{S}}$, where $X(t) = \mathbf{C}$ denotes that the queue vector at time t is $\mathbf{C}$. The Markov chain is irreducible, since transitions due to arrivals allow any state to be reached from the empty state and condition (ii) ensures that the empty state can be reached from any state by the departure transitions.

For $\mathbf{C} \in \widetilde{\mathcal{S}}$ and $c \in \mathcal{C}$ we associate an indicator function $I_c(\mathbf{C})$ such that $I_c(\mathbf{C}) = 1$ if $(c,\mathbf{C}) \in \widetilde{\mathcal{S}}$, and $I_c(\mathbf{C}) = 0$ if $(c,\mathbf{C}) \notin \widetilde{\mathcal{S}}$. Then for any $c \in \mathcal{C}$ and any $\mathbf{C} = (c_n,\ldots,c_1) \in \widetilde{\mathcal{S}}$, the transition rate due to type c arrivals is $\lambda_c r(n) I_c(\mathbf{C})$ and the transition rate due to the departure of a type c_i customer in queue position i when there are n customers in queue is $\mu(n)s_i(c_i,\ldots,c_1)$.

The functions $r(n)$ and $I_c(\mathbf{C})$ parametrise the rejection rule. Thus $r(n) = 0$ applies blocking when the queue length reaches some threshold n and $I_c(\mathbf{C}) = 0$ implements a more refined form of blocking which depends on how many customers of the various types are present.

The equilibrium equations for the Markov chain are given by

$$\pi(0) \sum_{c\in\mathcal{C}} \lambda_c r(0) I_c(0) = \mu(1)k(c) \sum_{(c)\in\widetilde{\mathcal{S}}} \pi(c) \tag{2.23}$$

and

$$\begin{aligned}\pi(\mathbf{C}) &\left(\sum_{c\in\mathcal{C}} \lambda_c r(n) I_c(\mathbf{C}) + \mu(n)k(\mathbf{C})\right) \\ &= \sum_{c\in\mathcal{C}} \sum_{i=0}^{n} \mu(n+1)\pi(c_n,\ldots,c_{i+1},c,c_i,\ldots,c_1)s_{i+1}(c,c_i,\ldots,c_1) \\ &\quad + \lambda_{c_n} r(n-1)\pi(c_{n-1},\ldots,c_1)\end{aligned} \tag{2.24}$$

for any state $\mathbf{C} = (c_n,\ldots,c_1) \in \widetilde{\mathcal{S}}$ with the convention that $\pi(\mathbf{C}) = 0$ if $\mathbf{C}$ is not in $\widetilde{\mathcal{S}}$.

In section 2.2 we showed that if $r(n) = 1$ for all n and $I_c(\mathbf{C}) = 1$ for all $c \in \mathcal{C}$ and $\mathbf{C} \in \widetilde{\mathcal{S}}$ (thus $\widetilde{\mathcal{S}} = \mathcal{S}$) then the solution to (2.23) and (2.24) can be found as a solution to the partial balance equations

$$\sum_{i=0}^{n} \frac{\pi(c_n,\ldots,c_{i+1}cc_i,\ldots,c_1)}{\pi(\mathbf{C})} \mu(n+1)s_{i+1}(cc_i,\ldots,c_1) = \lambda_c r(n) I_c(\mathbf{C}) \tag{2.25}$$

and

$$\mu(n)k(\mathbf{C})\pi(\mathbf{C}) = \lambda_{c_n} r(n-1)\pi(c_{n-1},\ldots,c_1) \tag{2.26}$$

for all $\mathbf{C} \in \widetilde{\mathcal{S}}$. This suggests the form of the stationary distribution for the OI loss queue which is stated in the following theorem.

Theorem 2.3. *If the service rate functions $s_i(\cdot)$ conform to the conditions (i) and (ii) on the set $\widetilde{\mathcal{S}}$ and $\widetilde{\mathcal{S}}$ conforms to conditions (a) and (b) then the stationary distribution of the Markov chain on $\widetilde{\mathcal{S}}$ associated with an OI loss queue is given by $\pi(0) = 1/G$ and*

$$\pi(c_n,\ldots,c_1) = \frac{1}{G}\prod_{i=1}^{n} \frac{\lambda_{c_i} r(i)}{\mu(i)k(c_i,\ldots,c_1)} \tag{2.27}$$

if and only if

$$G = 1 + \sum_{\substack{(c_n,\ldots,c_1)\in\widetilde{\mathcal{S}}\\ n>0}} \prod_{i=1}^{n} \frac{\lambda_{c_i} r(i)}{\mu(i)k(c_i,\ldots,c_1)} < \infty.$$

Further, the OI loss queue satisfies partial balance.

Proof. Equation (2.27) is clearly a solution to (2.26) on $\widetilde{\mathcal{S}}$. We need to show that (2.27) is also a solution to the partial balance equation (2.25) on $\widetilde{\mathcal{S}}$ and is thus a solution to the equilibrium equations. If the state $(c,\mathbf{C})$ does not belong to $\widetilde{\mathcal{S}}$ then from condition (b) $I_c(\mathbf{C}) = 0$ and from condition (a) any permutation of the state $(c,\mathbf{C})$ does not belong to $\widetilde{\mathcal{S}}$ implying that both sides of (2.25) are equal to zero and the equation is satisfied. If the state $(c,\mathbf{C})$ belongs to the set $\widetilde{\mathcal{S}}$ then $I_c(\mathbf{C}) = 1$ and (2.5) coincides with the form of (2.25) for an OI queue without losses [4] except for the arrival rate scaling factor $r(n)$ which does not affect the calculations and thus (2.27) is a solution to the equilibrium equations. □

Although most of the functions used in the definition of the OI loss queue are independent of the order of the customers in the queue, the queue order cannot be ignored. Indeed, let $\mathcal{N}$ be a set of non-negative integers and let $\mathcal{N}^C$ denote the $C-$fold product space of $\mathcal{N}$. Define a function A from $\widetilde{\mathcal{S}}$ into $\mathcal{N}^C$ such that $A(\mathbf{C}) = \mathbf{N} = (n_1,\ldots,n_C)$ whose c^{th} element n_c denotes the number of type c elements in the vector $\mathbf{C}$. Let $\widetilde{\mathcal{N}}$ denote the set of all vectors $\mathbf{N} \in \mathcal{N}^C$ for which there exists a vector $\mathbf{C} \in \widetilde{\mathcal{S}}$ such that $A(\mathbf{C}) = \mathbf{N}$. Consider the process $Y(t)$ on $\widetilde{\mathcal{N}}$ associated with the OI loss queue which records only the numbers of the customers of each type in the queue at time t. In general $Y(t)$ is not a Markov chain because the departure rates of the individual customer types $k_c(\mathbf{C}) = \sum_{i=1}^{n} s_i(c_i,\ldots,c_1)\,1(c_i = c)$ are not necessarily order independent.

If the $k_c(\mathbf{C})$ are order independent then $Y(t)$ is a Markov chain on the aggregated space $\widetilde{\mathcal{N}}$ and the following theorem holds.

Theorem 2.4. *Let $X(t)$ be a stationary Markov chain associated with an OI loss queue. If the service rate functions $k_c(\mathbf{C}), c \in \mathcal{C}$, are order independent on $\widetilde{\mathcal{S}}$ then the process $Y(t)$ is a reversible Markov chain on $\widetilde{\mathcal{N}}$ with a stationary distribution $\pi(\mathbf{N})$ which can be recursively calculated from the detailed balance equations*

$$\mu(n)k_c(\mathbf{N})\pi(\mathbf{N}) = \lambda_c r(n-1)\pi(\mathbf{N} - \mathbf{1}_c) \tag{2.28}$$

where $n_c > 0$, $n = n_1 + \cdots + n_C$ and $k_c(\mathbf{N}) = k_c(\mathbf{C})$ for $\mathbf{N} = A(\mathbf{C})$.

Proof. $Y(t)$ is a Markov chain on the aggregated space since all the transitions due to customer arrivals and departures are well defined. Thus the stationary distribution $\pi(\mathbf{N})$ on $\widetilde{\mathcal{N}}$ is

$$\pi(\mathbf{N}) = \sum_{\mathbf{C}:A(\mathbf{C})=\mathbf{N}} \pi(\mathbf{C}). \tag{2.29}$$

Equation (2.28) is obtained by summing (2.25) over all states $\mathbf{C}$ such that $A(\mathbf{C}) = \mathbf{N}$, taking the order independence of $k_c(\mathbf{C})$ into account. Equation (2.28) is the detailed balance equation for the Markov chain $Y(t)$ and is satisfied by the probabilities given by (2.29). Thus the Markov chain $Y(t)$ is reversible and its distribution can be recursively calculated from the detailed balance (2.28). □

Theorem 2.4 demonstrates that if the departure rates $k_c(\mathbf{C})$ of the individual customer types are order independent over the entire state space then the OI construction is not necessary to derive the equilibrium distribution and the queue can be examined via standard reversibility arguments as is shown for example in the model presented in section 2.5.7.

2.4.2 The Performance Measures: the MSCCC Loss Queue

Let $\mathbf{C} = (c_n, \dots, c_1)$ denote the state of the MSCCC loss queue where $n \le N$. The state space of the queue is $\mathcal{L} = \mathcal{S}_N = \{0\} \cup \bigcup_{n=1}^{N} \mathcal{C}^n$. Let $M_c(\mathbf{C})$ denote the number of type c customers in $\mathbf{C}$. Recall the mapping $\mathcal{A} : \mathcal{S}_N \mapsto \mathbf{Z}^C$ such that for any $\mathbf{C} \in \mathcal{S}_N$

$$\mathcal{A}(\mathbf{C}) = \mathbf{M}(\mathbf{C}) = (M_c(\mathbf{C}))_{c \in \mathcal{C}}$$

where $\mathbf{M}(\mathbf{C})$ is the counting vector of the state $\mathbf{C}$. Let $\mathcal{M}(N)$ denote the set of all counting vectors. As usual $|\mathbf{M}| = M_1 + \cdots + M_C$.

The stationary distribution $\pi_N(\mathbf{C})$ where $\mathbf{C} \in \mathcal{S}_N$ for the MSCCC loss queue will, on all states $\mathbf{C} \in \mathcal{S}_N$, coincide (up to a normalising constant) with the stationary distribution for the corresponding MSCCC queue with no losses. Define

$$\pi_N(\mathbf{M}) = \sum_{\substack{\mathbf{C} \in \mathcal{S}_N \\ \mathcal{A}(\mathbf{C}) = \mathbf{M}}} \pi_N(\mathbf{C}).$$

2.4.2.1 Invariant Measures over Special Sets

Recall that $k(\mathbf{M})$ denotes the number of customers in service when the MSCCC is in a state with a counting vector $\mathbf{M}$. For any $c \in \mathcal{C}$ and non-negative integers n, b, i define the sets

$$\begin{aligned}
\mathcal{L}(n) &= \{\mathbf{M} \in \mathcal{L} \mid |\mathbf{M}| = n\} \\
\mathcal{L}(n,b) &= \{\mathbf{M} \in \mathcal{L}(n) \mid k(\mathbf{M}) = b\} \\
\mathcal{L}(n,b,c) &= \{\mathbf{M} \in \mathcal{L}(n,b) \mid M_x = 0 \text{ if } x > c\} \\
\mathcal{L}(n,b,c,i) &= \{\mathbf{M} \in \mathcal{L}(n,b,c) \mid M_c = i\}.
\end{aligned}$$

In the remainder of this section, where no confusion can arise we omit the subscript N which denotes the population constraint on the MSCCC loss queue.

Let $Q(n)$ denote the probability that n customers are present at the MSCCC.

Let $Q(n,b)$ denote the probability that n customers are present and b servers are busy.

Let $Q(n,b,c)$ denote the probability that n customers are present, b servers are busy and no customers of types higher than c are present.

Let $Q(n,b,c,i)$ denote the probability that n customers are present, b servers are busy, i customers of type c are present and no customers of types higher than c are present.

The first step in the calculation of $Q(n)$ is to compute the $Q(n,b,c,i)$. In the domain $k(\mathbf{M}) < B$ where not all servers are busy, the number of type c customers in service is given by $k_c(\mathbf{M}) = M_c \wedge B_c$. For any $c \in \mathcal{C}$ and $0 < b < B$ two cases arise.

(1) $0 < i \leq n \leq N$

$$\begin{aligned}
Q(n,b,c,i) &= \sum_{\mathbf{M} \in \mathcal{L}(n,b,c,i)} Q(\mathbf{M}) \\
&= \sum_{\mathbf{M} \in \mathcal{L}(n,b,c,i)} \frac{\rho_c}{i \wedge B_c} Q(\mathbf{M} - \mathbf{1}_c) \\
&= \begin{cases} \dfrac{\rho_c}{i} Q(n-1, b-1, c, i-1) & 0 < i \leq B_c \\[2ex] \dfrac{\rho_c}{B_c} Q(n-1, b, c, i-1) & i > B_c \end{cases}
\end{aligned} \tag{2.30}$$

with the convention that $Q(n,b,c,i) = 0$ if the condition $0 < (i \wedge B_c) \leq b < B$ is violated.

(2) $i = 0$ and $0 \leq n \leq N$

$$\begin{aligned}
Q(n,b,c,0) &= \Pr(n \text{ customers present}, b \text{ servers busy}, M_x = 0 \text{ if } x \geq c) \\
&= \Pr(n \text{ customers present}, b \text{ servers busy}, M_x = 0 \text{ if } x > c-1) \\
&= Q(n,b,c-1).
\end{aligned} \tag{2.31}$$

The starting values for the recursion presented in (2.30) and (2.31) are given by

$$Q(n,0,c,0) = \begin{cases} 1 & n = c = 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.32}$$

for all $0 \leq n \leq N$ and $0 \leq c \leq C$. The next step is to compute the $Q(n,b,c)$

$$Q(n,b,c) = \sum_{i=0}^{n} Q(n,b,c,i) = Q(n,b,c-1) + \sum_{i=1}^{n} Q(n,b,c,i)$$

where

$$Q(n,0,c) = \begin{cases} 1 & n = c = 0 \\ 0 & \text{otherwise.} \end{cases} \tag{2.33}$$

The last step is to compute the $Q(n,b)$

$$\begin{aligned} Q(n,b) &= \Pr(n \text{ customers present and } b \text{ servers busy}) \\ &= \Pr(n \text{ customers present, } b \text{ servers busy and } M_x = 0 \text{ if } x > C) \\ &= Q(n,b,C) \end{aligned}$$

which yields

$$Q(N) = \sum_{b=0}^{B} Q(N,b). \tag{2.34}$$

Equation (2.34) cannot be used to compute $Q(N)$ since $Q(N,B)$ is not known. However, since the arrival rate of accepted traffic to the queue (the accepted traffic is the offered traffic minus the lost traffic) equals the departure rate from the queue

$$\rho(1 - Q(N)) = \sum_{b=0}^{B} bQ(b) = \sum_{b=0}^{B} b \sum_{n=b}^{N} Q(n,b) = \sum_{b=0}^{B} bR(N,b)$$

so that

$$BR(N,B) = \rho(1 - Q(N)) - \sum_{b=0}^{B-1} bR(N,b).$$

Adding $B\sum_{b=0}^{B-1} R(N,b)$ to both sides of the above equation yields

$$B = \rho(1 - Q(N)) + \sum_{b=0}^{B-1} (B-b)R(N,b) \tag{2.35}$$

so that

$$\rho Q(N) = \rho - B + \sum_{b=0}^{B-1} (B-b)R(N,b)$$

where $R(N,0) = \pi(\mathbf{0})$ yields an expression for $Q(N)$ in terms of the previously computed values of $Q(N,b)$ where $0 \le b < B$.

The $Q(N,b)$ are multiples of $\pi(\mathbf{0})$. To calculate $\pi(\mathbf{0})$, note that if $\overline{Q}(N)$ and $\overline{Q}(N,b)$ represent the values obtained by starting the algorithm with an arbitrary value of $Q(0,0,0,0)$ in (2.32), then

$$\rho(1 - \pi(\mathbf{0})\overline{Q}(N)) = \pi(\mathbf{0}) \sum_{b=0}^{B} b\overline{R}(N,b).$$

Subtracting $B\pi(\mathbf{0})\sum_{b=0}^{B}\overline{R}(N,b) = B$ from both sides of the above equation yields

$$B - \rho(1 - \pi(\mathbf{0})\overline{Q}(N)) = \pi(\mathbf{0})\sum_{b=0}^{B-1}(B-b)\overline{R}(N,b)$$

so that

$$\pi(\mathbf{0}) = \frac{B+\rho}{\sum_{b=0}^{B-1}(B-b)\overline{R}(N,b) - \overline{Q}(N)} \tag{2.36}$$

where $\overline{R}(N,0) = 1$ yields an expression for $\pi(\mathbf{0})$ in terms of the previously computed values of $Q(N,b)$ where $1 \leq b < B$. Then

$$Q(N,b) = \begin{cases} \pi(\mathbf{0})\overline{Q}(N,b) & 0 \leq b < B \\ \\ \pi(\mathbf{0})Q(N) - \sum_{b=0}^{B-1} Q(N,b) & b = B. \end{cases}$$

Note that $\lim_{N\to\infty} Q(N) = 0$ and $\lim_{N\to\infty} R(N,b) = P(b)$ so that in the limit where no losses occur, (2.35) and (2.36) reduce to (2.14) and (2.15).

Another recursion with the same computational complexity as the recursion presented in (2.30) and (2.31) but with O(B^2) as opposed to O(B^3) storage requirements is

$$Q(n,b,c) = \sum_{i=0}^{B_c-1}\frac{\rho_c^i}{i!}Q(n-i,b-i,c-1) + \frac{\rho_c^{B_c}}{B_c!}Q(n-B_c,b-B_c,c-1)F(n,c)$$

where $0 < c \leq C$ and $0 < b < B$ and

$$F(n,c) = \begin{cases} \dfrac{1-(\rho_c/B_c)^{n-B_c+1}}{1-\rho_c/B_c} & \rho_c \neq B_c \\ \\ n - B_c + 1 & \rho_c = B_c. \end{cases}$$

2.4.2.2 The Expected Queue Length

Let $L(n,b,C)$ denote the average number of type C customers at the MSCCC when $b < B$ servers are busy and n customers of all types are present. For $0 < b < B$

$$L(n,b,C) = \sum_{i=1}^{n} iQ(n,b,C,i)$$

where the $Q(n,b,c,i)$ are given in (2.30). Given the relatively small number n of terms in the above summation, it is not necessary as was in the case of the MSCCC without losses – see section 2.3.2.2 – to compute an analytic expression for the sum.

Let $L(n,C)$ denote the average number of type C customers at the MSCCC when there are n customers at the MSCCC

$$L(n,C) = \sum_{b=1}^{B} L(n,b,C).$$

The values of the $L(n,B,C)$ are not known. Consider therefore the following identity

$$BL(n,C) = \sum_{b=1}^{B} (B+b-b)L(n,b,C) = \sum_{b=1}^{B} bL(n,b,C) + \sum_{b=1}^{B-1} (B-b)L(n,b,C).$$

Since $k(\mathbf{M}) = b$ we have

$$\begin{aligned}
\sum_{b=1}^{B} bL(n,b,C) &= \sum_{b=1}^{B} b \sum_{\mathbf{M}\in\mathcal{L}(n,b)} M_C Q(\mathbf{M}) \\
&= \sum_{b=1}^{B} b \sum_{\mathbf{M}\in\mathcal{L}(n,b)} M_C \sum_{x=1}^{C} \frac{\rho_x}{k(\mathbf{M})} Q(\mathbf{M}-\mathbf{1}_x) \\
&= \sum_{x=1}^{C-1} \rho_x \sum_{b=1}^{B} \sum_{\mathbf{M}\in\mathcal{L}(n,b)} M_C Q(\mathbf{M}-\mathbf{1}_x) + \rho_C \sum_{b=1}^{B} \sum_{\mathbf{M}\in\mathcal{L}(n,b)} M_C Q(\mathbf{M}-\mathbf{1}_C) \\
&= \sum_{x=1}^{C-1} \rho_x \sum_{\mathbf{M}\in\mathcal{L}(n-1)} M_C Q(\mathbf{M}) + \rho_C \sum_{\mathbf{M}\in\mathcal{L}(n-1)} (M_C+1) Q(\mathbf{M}) \\
&= \rho L(n-1,C) + \rho_C Q(n-1)
\end{aligned}$$

where $\rho = \sum_{c=1}^{C} \rho_c$ so that

$$BL(n,C) = \rho L(n-1,C) + \rho_C Q(n-1)) + \sum_{b=1}^{B-1} (B-b)L(n,b,C).$$

The last term $L(N,C)$ of this recursion is the expected number of type C customers in the MSCCC loss queue. Summing the values of $L(N,c)$ over all $c \in \mathcal{C}$ yields the expected queue length – which result can be obtained at earlier stage from $\sum_{n=1}^{N} nQ(n)$.

2.4.3 OI Loss Networks

Although we have presented the OI system as a single queue, OI systems can also be used to model networks of queues.

Let $\mathbf{C}(t)$ be a continuous time, homogeneous Markov process denoting the state of an OI queue at time t on the state space $\widetilde{\mathcal{S}}$ with the stationary distribution $\pi(\mathbf{C})$. Let $\widetilde{\mathcal{S}'}$ be a subset of $\widetilde{\mathcal{S}}$ satisfying conditions (a) and (b). Then by definition $\mathbf{C}(t)$ is an OI loss queue on the set $\widetilde{\mathcal{S}'}$ with a stationary distribution given by $\pi(\mathbf{C})/G$ where G is the corresponding normalising constant. This truncation property of the OI queue can be extended to a network of independent OI queues as follows.

Let $\mathbf{C}^k(t) = (c^k_{n_k}(t),\ldots,c^k_1(t))$ where $k = 1,\ldots,K$ denote a collection of independent OI queues on the corresponding states $\widetilde{\mathbb{S}}^k$ with the stationary distribution $\pi^k(\mathbf{C}^k)$. Consider the queue $\mathbf{C}(t) = (\mathbf{C}^1(t),\ldots vC^K(t))$ on the state $\widetilde{\mathbb{S}} = \widetilde{\mathbb{S}}^1 \times \widetilde{\mathbb{S}}^2 \times \cdots \times \widetilde{\mathbb{S}}^K$ with stationary distribution $\pi(\mathbf{C}) = \pi^1(\mathbf{C}^1)\ldots\pi^K(\mathbf{C}^K)$. Let $\widetilde{\mathbb{S}}'$ be a subset of $\widetilde{\mathbb{S}}$ such that

(i) $\mathbf{C} \in \widetilde{\mathbb{S}}'$ implies $(\mathbf{C}^1_\sigma,\ldots,\mathbf{C}^K_\sigma) \in \widetilde{\mathbb{S}}'$ for any permutations $\mathbf{C}^k_\sigma$ of $\mathbf{C}^k$.
(ii) $A(\widetilde{\mathbb{S}}') = \widetilde{\mathbb{N}}'$ is coordinately convex

then [5] the queue $\mathbf{C}(t)$ obeys the truncation property so that the stationary distribution of $\mathbf{C}(t)$ on the set $\widetilde{\mathbb{S}}'$ is given by $\pi(\mathbf{C})/G$ where G is the corresponding normalising constant.

Several applications of OI queues and OI networks are presented in the following section.

2.5 OI Applications

2.5.1 Multiported Memory

Consider [7, 8, 15] a computer system consisting of N processors accessing K memory modules via a partitioned multiple bus system. Each of the G groups of B buses gives access to a subset of K/G memory modules. Each memory module k is B_k–ported so that maximally B_k processors can access memory module k simultaneously. The system is modelled as a closed queueing network consisting of an IS centre representing the processors and G MSCCC centres, each representing one group of B buses and the associated memory modules. Each MSCCC centre consists of B servers which represent the B buses in its group. The N customers in the network belong to K classes.

A processor service interval followed by a data transfer to/from memory module k is modelled as a customer departing from the processor service centre, and moving to the g^{th} MSCCC centre where group g contains an access path to memory module k. The customer changes class to class k and queues for service at the MSCCC centre. The class k customer enters into service if one of the B servers is free (a bus is available) and if at most B_k class k customers are in service (at most B_k processors can access memory module k simultaneously).

2.5.2 A Messaging Card

In some distributed architectures such as telephone switching exchanges, the messaging function between a high level peripheral (decentralised call processor) and a lower level one (line or trunk controller) is performed on the high level side by a specialised processor (the messaging card) which controls simplex channels to the lower level peripherals. The processor time is partitioned into B fixed time slots, each of which is allocated to a process whose function is to send outgoing messages. The exchange forwards messages to K destinations, each of which is reachable on any of B_k channels. When a buffer is queued for transmission it has to wait for a process to be available to service the request and for an outgoing channel to be free. The messaging card can be modelled by a MSCCC centre consisting of B servers (the transmission processes) serving customers belonging to K classes, each with its own concurrency limit B_k.

2.5.3 Multilayer Window Flow Control

Consider a set of K application which share a common data link. At the data link layer maximally B transmitted packets may remain unacknowledged. For each application k maximally B_k packets may remain unacknowledged. The data link can be modelled [12] as a MSCCC centre consisting of B servers with an average service time equal to the packet transfer time averaged over all the packets. The customer arrivals represent packet transmission requests. A class k packet is transmitted if a server is available (data link flow control) and if at most $B_k - 1$ class k packets are in service (application flow control).

2.5.4 Machine Scheduling Model

Jobs of type $k \in \{1,\dots,C\}$ arrive according to a Poisson process with intensity λ_k and are processed by machines of type k in a FCFS order. There are B_k machines of type k which are operated by a common pool of B machine operators. If the job processing times are exponentially distributed and are job type independent then this model is solved by the MSCCC centre. If the job processing times are exponentially distributed but are job type dependent, then in the two special cases $B = 1$ and $B = B_1 + \cdots + B_C$ the stationary probabilities are a sum of product forms [1].

2.5.5 Blocked Calls Cleared

In the remainder of this section we apply OI queues to model circuit-switched networks where some blocked calls are queued for connection when the requested circuits become available. This call queueing mechanism differs from the admission control generally used in circuit switched networks where blocked calls are lost – see [19, 20] for an extensive literature on the subject of circuit-switched loss networks.

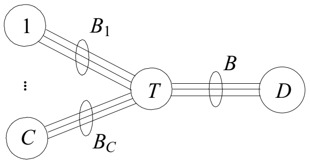


Fig. 2.1: A circuit-switched network

Consider figure 2.1 which represents (part of) a circuit-switched network. There are C source switches. Each switch c is connected by an access link consisting of B_c circuits to a tandem switch T which switches the incoming calls among B outgoing circuits to a destination switch D where $B < B_1 + \cdots + B_C$. Calls arrive at switch c according to a Poisson process with rate λ_c. Call holding times are exponentially distributed with unit mean.

A call from switch c is connected if a circuit from c to T and a circuit from T to D is available, else the call is lost. This model is OI although standard methods can also be used [23] to obtain an analytic solution. Such loss models were considered in a more general framework [2, 19] and there are many articles on this topic for example [9].

2.5.6 Blocked Calls Queued

Rather than clearing blocked calls, the blocked calls can be held for a short period while waiting for the required circuits to become free. This will significantly decrease the loss probability and increase the circuit utilisation at the expense of introducing a small connection delay.

Call queueing is implemented by storing the signalling information for each call in a buffer at the transit switch T until the call is completed. When a call completes, the queue is scanned from the front looking for the first call that can be connected. If the call (say it originated at switch c) cannot be connected because all B_c circuits in link c are busy, the next call in the queue is considered. Connection requests are thus attempted on a FCFS basis.

The MSCCC queue provides an exact analytic solution to this model. Calls which arrive and find N calls in the system (queued or in service) are lost. A call from switch c will be connected to the destination switch D if a circuit in link c is available (less than B_c customers of type c are in service) and a circuit from T to D is available (less than B customers of all types are in service). The limit N models call holding: up to $N - \min(B_1, \ldots, B_C)$ blocked calls can be queued. When $N = B$ we obtain Mitra's model [23].

A feature of this model is that if the network is overloaded by class c calls then the queue will be filled with class c calls and arrivals of other classes will be rejected. This deficiency is addressed in the next section.

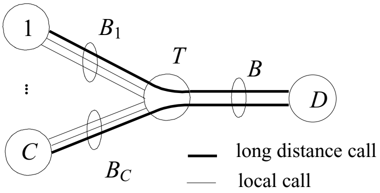


Fig. 2.2: Local and long distance calls

2.5.7 Blocked Calls Queued with Source Rejection

Calls which arrive and find N calls in the system (queued or in service) are lost. A call from switch c which finds all B_c access circuits busy is lost – this is termed source rejection. Calls which seize an access circuit and find all B outgoing circuits busy are queued in a finite buffer. When an outgoing circuit becomes free, the queued calls are searched in FCFS order for the next call to be connected. Thus maximally N customers of all types and maximally B_c customers of type $c \in \mathcal{C}$ are admitted to the system and up to $N - \min(B_1, \dots, B_C)$ blocked calls can be queued.

For all $\mathbf{n} \in \widetilde{\mathcal{N}}$ the un-normalised aggregated stationary probabilities for above model are given by

$$\pi(\mathbf{n}) = g(n) \prod_{c=1}^{C} \frac{\rho_c^{n_c}}{n_c!}$$

where

$$g(n) = \begin{cases} 1 & 0 \le n \le B \\ \dfrac{n!}{B!B^{n-B}} & B < n \le N \end{cases}$$

so that $Bg(n) = ng(n-1)$ for $B < n \le N$. This model is an OI queue although standard methods can also be used [6] to obtain an analytic solution.

A further OI specialisation of this model is obtained by partitioning the set of customer types into $\{1, \dots, J\}$ and $\{J+1, \dots, C\}$. Source rejection is applied to the types $\{J+1, \dots, C\}$ but is not applied to the types $\{1, \dots, J\}$ so that all customers of types $\{J+1, \dots, C\}$ are in service whereas customers of types $\{1, \dots, J\}$ may be queued.

2.5.8 Local and Long Distance Calls

Consider the network presented in figure 2.2. A local call from switch c requires one circuit from switch c to the tandem switch T. A long distance call from switch c requires one circuit from switch c to switch T and one circuit from T to the destination switch D. Blocked local calls are lost. Blocked long distance calls are queued in a finite buffer of size $N - \min(B_1, \dots, C)$.

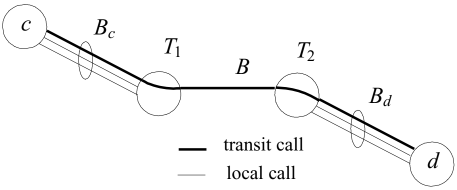


Fig. 2.3: Local and transit calls

Let $\mathcal{C}_+ \subseteq \mathcal{C}$ denote the set of long distance call types and let $n_+ = \sum_{j\in\mathcal{C}_+} n_j$ denote the number of long distance calls, both in service and queued. The unnormalised aggregated stationary probability for the queue is given by [5]

$$P(\mathbf{n}) = \begin{cases} \displaystyle\prod_{j\in\mathcal{C}} \frac{\rho_j^{n_j}}{n_j!} & 0 \le n_+ \le B \\ \displaystyle\frac{n_+!}{B!B^{n_+-B}} \prod_{j\in\mathcal{C}} \frac{\rho_j^{n_j}}{n_j!} & B < n_+ \le N \end{cases}$$

where $\rho_j = \lambda_j/\mu_j$. Note that $P(\mathbf{n})$ can be viewed as the product of two "independent" queues: an $M/M/B/B$ queue serving the local calls and an $M/M/B/N$ queue serving the long distance calls. The processing of the local and long distance calls is not independent.

2.5.9 Local and Transit Calls

Consider the network presented in figure 2.3. A local call from switch c requires one circuit from link c. A local call from switch d requires one circuit from link d. Blocked local calls are lost. A transit call from switch c to switch d requires one circuit from link c, one circuit from link d and a circuit from switch T_1 to switch T_2. A transit call is lost if it is blocked on link c or on link d. If the transit call is not blocked on the local links it seizes the local circuits and requests a circuit from T_1 to T_2. If the circuit is available it is connected. If the circuit is not available and if there are less than N transit calls in progress (queued and in service) on the link from T_1 to T_2 then the transit call is queued, else it it is lost.

2.5.10 Hierarchical Tree Networks

Several networks as presented in figure 2.1 are connected to form the network shown in figure 2.4. A call from switch c is lost if it is blocked on any link connecting the source switch c to the switch T. A finite queue is used to buffer calls that are blocked on the final link from T to D. This model is a variant of the Multiserver Centre with Hierarchical Classes of Customers [21].

2.5.11 Local and External Networks

Consider the network presented in figure 2.5. The network consists of edge switches marked E and and interior switches marked I in a core network and access switches marked A in an access network. The switches are connected by multi-circuit links. The topology of the network is arbitrary and we assume that the switches in the access network are not directly connected to each other, although this assumption is not necessary.

Local calls are routed among the switches in the core network. Blocked local calls are lost. Consider an outgoing call from the network via an edge switch E to an access switch A. An outgoing call is lost if it is blocked on its local route. If the outgoing call is not blocked on its local route then it seizes the circuits in its local route and requests a circuit in the outgoing link from switch E to switch A. If an outgoing circuit is available the call is connected. If an outgoing circuit is not available and if there are less than N outgoing calls in progress (queued and in service) on the outgoing link then the outgoing call is queued, else it it is lost. Note that the local circuits acquired by outgoing calls are held while the calls are queued for connection, and that each outgoing link carries traffic in the outgoing direction only.

Call queueing is accomplished by storing the signalling information for the blocked call in a buffer at the edge switch E. The (signalling units for the) calls

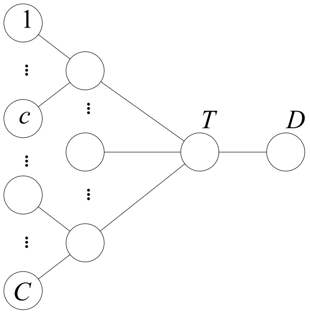
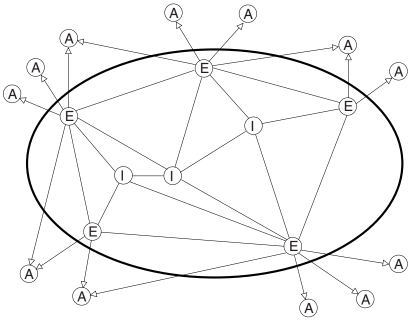


Fig. 2.4: Hierarchical tree network



A access switches E edge switches I interior switches

Fig. 2.5: Network with outgoing traffic

are stored in FCFS order. When a call completes and releases a circuit in the outgoing link to the access switch A the first call in the queue will be connected to the destination A.

Let γ denote a fixed route (an ordered sequence of links) connecting an originating switch I in the network via an edge switch E to an access switch A. Let $\mathcal{R}$ denote the set of interior routes. Let n_γ denote the number of calls in route $\gamma \in \mathcal{R}$ and let $\mathbf{n} = (n_\gamma)$ be a vector of numbers of calls. Let n_k where $k \in \mathcal{K}$ denote the number of calls (queued or in service) in progress on outgoing link k where $\mathcal{K}$ is the set of access links, and $\mathcal{R} \cap \mathcal{K} = \emptyset$. Then the un-normalised aggregated stationary probabilities for the network are given by [5]

$$P(\mathbf{n}) = \prod_{k \in \mathcal{K}} g_k(n_k) \prod_{\gamma \in \mathcal{R}} \frac{\rho_\gamma^{n_\gamma}}{n_\gamma!}$$

and

$$g_k(n_k) = \begin{cases} 1 & 0 \le n_k \le B_k \\ \dfrac{n_k!}{B_k! B_k^{n_k - B_k}} & B_k < n_k \le N_k \end{cases}$$

2.5.12 Transit Calls among Networks

Consider the system presented in figure 2.6. Each network consists of switches connected by multi-circuit links. Some of the switches are connected to gateway switches which route transit traffic in both directions between the networks. The topology of the networks is arbitrary. We assume one pair of gateway switches but there may be many such pairs.

Local calls are routed among switches within their originating networks. Blocked local calls are lost. Consider a transit call from network 1 to network 2. A transit call is lost if it is blocked on its local route in network 1 from its originating switch to the gateway switch G or if it is blocked on its local route in network 2 from the gateway switch G to its destination switch (assuming a signalling link separate from the bearer link). If the transit call is not blocked on its local routes then it seizes the circuits in its local routes and requests a circuit on the transit link between the networks. If a transit circuit is available the call is connected. If a transit circuit is not available and if there are less than N transit calls in progress (queued and in service) on the transit link then the transit call is queued, else it it is lost.

The transit link carries two-way traffic and an identical call queueing mechanism is applied to calls in both directions. Each gateway has a buffer to store the signalling units for blocked calls. Both gateway switches are aware via the signalling link of the arrival order of blocked calls in both directions. The buffers are managed as a single logical queue whose entries are ordered according to the times of arrival of the blocked calls. The call resources are partitioned into two classes (local routes and transit routes) and the resource classes are ordered such that all calls first request their local routes and then their transit route – deadlock therefore cannot occur.

All of the models presented in sections 2.5.1 through 2.5.12 can be modelled as OI queues. Their stationary distributions are therefore immediately available. Efficient algorithms with space-time complexity $\mathrm{O}(B^2C)$ exist for the calculation of the blocking probabilities for models 2.5.5, 2.5.6 and 2.5.7. The complexity of model 2.5.10 is $\mathrm{O}(N^\ell)$ where ℓ is the depth of the tree. Models 2.5.9, 2.5.11 and 2.5.12 have the same complexity as the corresponding loss networks where blocked calls are lost.

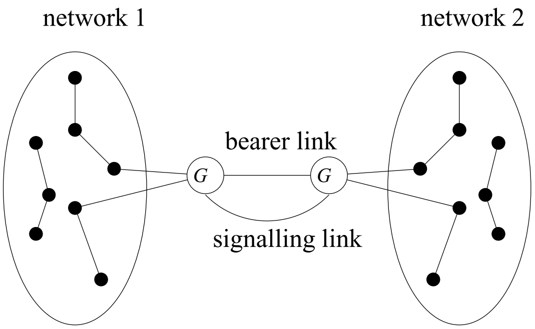


Fig. 2.6: Networks with transit traffic

Glossary

B	the number of servers
B_c	the number of type c customers that can be in service simultaneously
$\mathcal{C}$	the set of customer types
$\mathbf{C}$	the queue state $(c_n,\ldots,c_1)$
c_i	the type of the customer in queue position i
$\gamma_i(\mathbf{C})$	the portion of the total service effort directed at the customer in queue position i
λ_c	the Poisson arrival rate of customers of type c
μ_c	the Poisson service rate of customers of type c
$\pi(\mathbf{C})$	the equilibrium probability
$\phi(\mathbf{C})$	the rate at which the total service effort in state $\mathbf{C}$ is supplied
$\mathcal{S}$	the state space of the queue
$s_i(\mathbf{C})$	the rate at which service is given to the customer in queue position i in the queue relative to the other customers in the queue
$a \wedge b$	the smaller of the two integers a and b
σ	a permutation of $(1,\ldots,n)$

2.6 An Algorithm to Compute the Performance Measures of the MSCCC

The values of the $P(b,c,i)$ are stored in the elements of the array $P[b,i]$. The index c is suppressed to save storage. The algorithm is applied C times to compute the performance measures of all the types. After each application the customer types are relabelled.

```
// allocate and initialise the variables
double $P[0:B-1,0:\max(B_1,\ldots,B_C)],L[0:B-1]$
P[0,0] = 1
**for** $b=1$ to $B-1$ **do**
    $P[b,0]=0$
**end for**
// compute the un-normalised probabilities $P(b,c,i)$
**for** $c=1$ to $C$ **do**
    **for** $b=1$ to $B-1$ **do**
        **for** $i=1$ to $\min(b,B_c-1)$ **do**
            $P[b,i]=\rho_c * P[b-1,i-1]/i$                    // eq. (2.9)
        **end for**
        **if** $b \geq B_c$ **then**
            $P[b,B_c]=\rho_c * P[b-1,B_c-1]/(B_c-\rho_c)$     // eq. (2.10)
        **end if**
    **end for**
    **for** $b=1$ to $B-1$ **do**
```

```
        for $i = 1$ to $\min(b, B_c)$ do
            $P[b,0] = P[b,0] + P[b,i]$                         // eq. (2.13)
        end for
    end for
end for
// normalise the probabilities $P(b,c,i)$
$S = 0$
for $b = 0$ to $B-1$ do
    $S = S + (B-b) * P[b,0]$                                   // eq. (2.15)
end for
$G = (B-\rho)/S$
for $b = 0$ to $B-1$ do
    for $i = 1$ to $\min(b, B_c - 1)$ do
        $P[b,i] = P[b,i] * G$                                  // eq. (2.16)
    end for
end for
// compute the type $C$ queue length $L(C)$
$L = \rho_C$
for $b = 1$ to $B-1$ do
    $L[b] = P[b, B_C]$
    for $i = 0$ to $\min(b, B_C - 1)$ do
        $L[b] = L[b] + B_C * P[b-1,i] - i * P[b,i]$            // eq. (2.19)
    end for
    $L = L + (B-b) * L[b]$                                     // eq. (2.22)
end for
$L = L * \rho_C/(B_C - \rho_C)/(B-\rho)$                       // eq. (2.22)
```

References

1. I. Adan, J. Visschers and J. Wessels. Sum of product forms solutions to MSCCC queues with job type dependent processing times. Memorandum COSOR 98-19, Eindhoven University of Technology, The Netherlands (1998).
2. J.M. Akinpelu, The Overload Performance of Engineered Networks With Nonhierarchical and Hierarchical Routing, International Teletraffic Congress, Vol 10 (1983) 3.2.4.1 – 3.2.4.7.
3. F. Baskett, K.M. Chandy, R.R. Muntz and J. Palacios, Open, Closed and Mixed Networks of Queues with Different Classes of Customers, Journal of the ACM Vol 22 No 2 (1975) 249 – 260.
4. S.A. Berezner and A.E. Krzesinski, Quasi-reversible multiclass queues with order independent departure rates. Queueing Systems, 19 (1995) 345 – 359.
5. S.A. Berezner and A.E. Krzesinski, Order Independent Loss Queues. Queueing Systems, 23 (1996) 331 – 335.
6. S.A. Berezner and A.E. Krzesinski, Call Queueing in Circuit-Switched Networks. Telecommunication Systems Vol 6 (1996) 147 – 160.
7. J.-Y. Le Boudec, A BCMP Extension to Multiserver Stations with Concurrent Classes of Customers. In: *Proc. 1986 ACM Sigmetrics Conference, Performance Evaluation Review* Vol 14 No 1 (1986) 78 – 91.

8. J.-Y. Le Boudec, The MULTIBUS Algorithm. Performance Evaluation Vol 8 No 1 (Feb 1988) 1 – 18.
9. D.Y. Burman, J.P. Lehoczky and Y. Lim, Insensitivity of blocking Probabilities in a Circuit-Switching Network, Journal of Applied Probability, Vol 21 (1984) 850 – 859.
10. X. Chao and M. Pinedo, On Generalized Networks of Queues with Positive and Negative Arrivals, Prob. Eng. Inf. Sci, Vol 7 (1993) 301 – 334.
11. S. Crosby and A.E. Krzesinski, Product Form Solutions for Multiserver Centers with Concurrent Classes of Customers, Performance Evaluation, Vol 11 No 4 (1990) 265 – 281.
12. S. Crosby, A.E. Krzesinski and J.-Y. Le Boudec, A MSCCC Model of Multilayer Window Flow Control. In: *Fifth International Conference on Modeling Techniques and Tools for Computer Performance Evaluation*, G Balbo and G. Serazzi, Eds. (Elsevier 1992).
13. J.M. Harrison and R.J. Williams, On the Quasi-Reversibility of a Multiclass Brownian Station, Annals of Probability Vol 18 (1990) 1249 – 1268.
14. W. Henderson, C.E.M. Pearce, P.K. Pollett, P.G. Taylor, Connecting Internally Balanced Quasi-Reversible Markov Processes, Advances in Applied Probability Vol 24 (1992) 934 – 959.
15. M. Hofri and Y. Kogan, Exact and Asymptotic Analysis of Large Multiple-Bus Multiprocessor Systems, Proc. Performance ’90, Elsevier Science Publishers B.V. (North Holland) (1990) 373–389.
16. J.R. Jackson, Networks of Waiting Lines, Operation Research Vol 5 (1957) 518 – 521.
17. F.P. Kelly, Reversibility and Stochastic Networks, John Wiley and Sons (1979) ISBN 0-471-27601-4.
18. F.P. Kelly, Networks of Quasi-reversible Nodes, Applied Probability – Computer Science, the Interface: Proceedings of the ORSA–TIMS Boca Raton Symposium, Ed. R. Disney, Birkhauser Boston, Cambridge, Ma (1981).
19. F.P. Kelly, Blocking probabilities in Large Circuit-switching Networks, Advances in Applied Probability Vol 18 (1986) 473 – 505.
20. F. P. Kelly, Loss Networks, Annals of Applied Probability, Vol. 1, No. 3, (1991) 473 – 505.
21. A.E. Krzesinski and R. Schassberger, The Multiserver Center with Hierarchical Concurrency Constraints, Prob. Eng. Inf. Sci, Vol 6 (1992) 147 – 156.
22. M.A. Marsan, G. Balbo, G. Chiola and S. Donatelli, On the Product Form Solution of a Class of Multiple-Bus Multiprocessor System Models, Journal of Systems and Software Vol 1 No 2 (1986) 117 – 124.
23. D. Mitra, Asymptotic Analysis and Computational Methods for a Class of Simple Circuit-Switched Networks with Blocking, Advances in Applied Probability, Vol 19 (1987) 219 – 239.
24. R.R. Muntz, Poisson Departure Processes and Queueing Networks, IBM Research Report RC4145, IBM Thomas J. Watson Research Center, Yorktown Heights, New York (1972).
25. R.D. Nelson, The Mathematics of Product Form Queueing Networks, ACM Computing Surveys, Vol 25 (1993) 339 – 369.
26. E. Pinsky and A. Conway, Exact computation of blocking probabilities in state-dependent multi-facility blocking models, in *IFIP WG 7.3 International Conference on the Performance of Distributed Systems and Integrated Communication Networks*, T. Hasegawa, H. Takagi and Y. Takahashi eds., Kyoto, Japan 1991.
27. P.K. Pollett, Connecting Reversible Markov Processes, Advances in Applied Probability, Vol 18 (1986) 880 – 909.
28. P.K. Pollett, Preserving Partial Balance in Continuous-time Markov Chains, Advances in Applied Probability Vol 19 (1987) 431 – 453.
29. J. Walrand and P. Varaiya, Interconnections of Markov Chains and Quasi-Reversible Queueing Networks, Stochastic Processes and Applications Vol 10 (1980) 209 – 219.
30. J. Walrand, An Introduction to Queueing Networks, Prentice-Hall (1988), ISBN 0-13-493818-6.
31. J. Walrand, A Probabilistic Look at Networks of Quasireversible Queues, IEEE Trans. Inf. Theory IT-29 (1983) 825 – 831.
32. P. Whittle, Systems in Stochastic Equilibrium, John Wiley and Sons (1986) ISBN 0-471-90877-8.

Chapter 3
Insensitivity in Stochastic Models

P.G. Taylor

Abstract A stochastic model is said to be *insensitive* if its stationary distribution depends on one or more of its constituent lifetime distributions only through the mean. Insensitivity is usually associated with partial balance in the corresponding Markovian model when all lifetimes are taken to be exponential, and a product-form stationary distribution of the Markov chain, constructed by supplementing the state by information on the progress of generally-distributed lifetimes.

In this chapter I shall discuss insensitivity by presenting a detailed analysis of the canonical insensitive queueing model, the Erlang loss system, from two different directions, as a queue and as a Generalised Semi-Markov Process (GSMP). I shall then show how the underlying ideas extend to insensitive queueing network models and finish off with a discussion of the few known non-standard insensitive systems which are not associated with partial balance or a product-form supplemented stationary distribution.

3.1 Introduction

We shall start our discussion of insensitivity by thinking about the M/M/C/C (or Erlang Loss) queue. This is a queueing system which has Poisson arrivals, exponential service times, C servers and no room for queueing customers that arrive when the system is full. The queue can be modelled by a continuous-time Markov chain with state space $\{0,1,2,\dots,C\}$. If we denote the arrival rate by λ and the mean service time by $1/\mu$, then the stationary probability $\pi(n)$ that there are n customers present satisfies the equations

P.G. Taylor
Department of Mathematics and Statistics, University of Melbourne, Victoria, 3010, Australia
e-mail: p.taylor@ms.unimelb.edu.au

R.J. Boucherie and N.M. van Dijk (eds.), *Queueing Networks: A Fundamental Approach*,
International Series in Operations Research & Management Science 154,
DOI 10.1007/978-1-4419-6472-4_3, © Springer Science+Business Media, LLC 2011

$$\begin{aligned}
\lambda\pi(0) &= \mu\pi(1), \\
(\lambda+n\mu)\pi(n) &= \lambda\pi(n-1)+(n+1)\mu\pi(n+1), \quad 0<n<C, \\
C\mu\pi(C) &= \lambda\pi(C-1).
\end{aligned} \tag{3.1}$$

The solution of equations (3.1) that sums to unity is

$$\pi(n) = \frac{\rho^n/n!}{\sum_{k=0}^{C}\rho^k/k!}, \tag{3.2}$$

where $\rho = \lambda/\mu$. The stationary probability

$$\pi(C) = \frac{\rho^C/C!}{\sum_{k=0}^{C}\rho^k/k!} \tag{3.3}$$

that the system is full gives the probability that arriving customers cannot be accommodated in the queue.

Expressed as a function of ρ and C, the expression on the right hand side of equation (3.3) is known as Erlang's Loss Formula, which we shall denote by $E(\rho,C)$. Throughout most of the twentieth century, this formula was used extensively by the telecommunications networking community for dimensioning links. It proved to be remarkably successful in predicting the probability that an arriving call would not be able to find an available circuit, has been the subject of research in its own right (see, for example, Jagerman [22]) and is still used today in more complicated contexts.

However, let us think a little more about the use of a Markovian model for the modelling of telephone links. The average duration of a traditional phone conversation was three minutes. An easy calculation shows that if call durations are exponentially distributed with mean three minutes, then the probability that a call exceeds 60 minutes is about 2×10^{-9}. So, if call durations really were exponentially distributed, very few of us would ever have made a phone call that lasted longer than one hour. Since most of us have made such calls, we are led to the conclusion that the 'service times' corresponding to real telephone conversations are not exponentially distributed and that a Markovian model for the system is based upon assumptions that are not satisfied.

So why has the Erlang Loss Formula been so successful? The reason is that the M/G/C/C queue is *insensitive* to the service time distribution: the stationary probability that there are n customers present is given by (3.2) irrespective of the shape of the service time distribution, provided that the mean is $1/\mu$.

Erlang himself [10] noticed that the stationary probability that there are n customers present in an M/G/C/C queue when the service times are deterministic with duration $1/\mu$ is the same as it is when service times are exponentially distributed with mean $1/\mu$. Subsequently, with different levels of rigour, Kosten [33], Fortet [12] and Sevastyanov [43] showed that the service time distribution can be arbitrary without affecting the form of the stationary probabilities, assuming that the mean is kept constant. In fact more is possible: service times can be inter-event times in an arbitrary stationary point process with rate μ and the stationary distribution is still given by (3.2), see König and Matthes [31]. Other authors who considered the n server loss system with generally distributed service times from the point of view of insensitivity include Takacs [44] who investigated the stationary distribution at arrival epochs and Fakinos [11] who looked at a group arrival, group departure system.

In the period between the late 1950s and the 1980s a number of researchers studied the phenomenon of insensitivity in other systems. The Engset loss system, which has a finite source population, was shown to be insensitive with respect to generally distributed, but independent, service times by Cohen [9] and with respect to successive service times which come from a stationary point process by König [29]. More significant from a practical point view was the work of Baskett, Chandy, Muntz and Palacios [4] and Kelly [27, 28] who showed that certain types of queueing network possess the insensitivity property. Since a number of practical systems turned out to be well-modelled by insensitive queueing networks, these papers have been frequently cited, particularly in the telecommunications modelling community.

Baskett, Chandy, Muntz and Palacios [4] considered a network of queues where each node could be one of four different types. These were

1. a single server, first-come-first-served queue with exponential service times,
2. a single server, processor-sharing queue with service times chosen according to a general distribution with a rational Laplace Transform,
3. an infinite-server queue with service times chosen according to a general distribution with a rational Laplace Transform, and
4. a single-server, preemptive-resume last-come-first-served queue with service times again chosen according to a general distribution with a rational Laplace Transform.

They showed that the queueing network possesses a steady state distribution that is a product form over the nodes and, moreover, depends on the lifetime distribution at types (2), (3) and (4) nodes only through the mean. Weak continuity arguments [1, 47] later showed that the restriction to distributions with rational Laplace transform was unnecessary, although many later papers continued to emphasize this restriction.

Kelly [27, 28] introduced the concept of the *symmetric* queue. This can be thought of as a generalisation of the type (2), (3) and (4) nodes of [4]. A symmetric queue is a queue with multiple customer classes that operates in the following manner:

1. the service requirement of a customer is a random variable whose distribution may depend on the class of customer.
2. the total service effort is supplied at rate $\phi(n)$ where n is the number of customers in the queue.
3. a proportion $\gamma(\ell,n)$ of this effort is directed to the customer in position ℓ. When this customer leaves the queue customers in positions $\ell+1,\ell+2,\ldots,n$ move to positions $\ell,\ell+1,\ldots,n-1$ respectively.

4. a customer arriving at the queue moves into position ℓ with probability $\gamma(\ell, n+1)$. Customers previously in positions $\ell, \ell+1, \ldots, n$ move to positions $\ell+1, \ell+2, \ldots, n+1$ respectively.

Processor sharing queues, infinite server queues and last come first served queues are all examples of symmetric queues. By keeping track of the current 'phase' of service, Kelly showed that a stationary symmetric queue is insensitive to the service time distribution, provided that it can be represented as a mixture of Erlang distributions. The rigorous extension to arbitrarily distributed lifetimes was carried out by Barbour [1]. Furthermore, Kelly established that a network of symmetric queues has a stationary distribution that factorizes into a product form over the nodes, and itself is insensitive.

Kelly's techniques relied on the insight provided by the time-reversed process. Chandy, Howard and Towsley [7] used a partial balance approach to prove insensitivity in an essentially similar system. Noetzel [37] defined Last Batch Processor Sharing (LBPS) disciplines for queues, and showed that networks of LBPS queues have product-form stationary distribution and are insensitive. Noetzel focused on the arrival order, rather than the position in the queue, of customers but it is possible to set up an equivalence between LBPS queues and symmetric queues and so derive Noetzel's results from Kelly's.

Jansen and König [25] modified Chandy, Howard and Towsley's network to include different classes of customer and showed that, when the network is insensitive, the output processes from nodes in a queueing network are Poisson. They also considered the stationary distribution embedded at jump epochs of the system. Further work on queueing networks with multiple customer classes was given in Chandy and Martin [8].

Hordijk and van Dijk [17] studied networks of queues with blocking. They showed that there is some trade off between the generality of the blocking function and the degree of balance required from the routing matrix for a network to have product-form stationary distribution and possess an insensitivity property. In [18, 19], they introduced a new method for analysing networks of queues that depends on an associated process called the adjoint process. Using this approach they showed that a queue must be symmetric (in the definition of Kelly) to satisfy their concept of job local balance. They also applied their analysis to a range of models with general routing and service characteristics.

A general framework for studying insensitivity, the Generalised Semi-Markov Process (GSMP), was introduced by Matthes [34]. The GSMP is basically an extension of the familiar Semi-Markov Process, which resides in a particular state for a generally distributed length of time before undergoing a transition according to a stochastic matrix which transfers the process to another state. In a GSMP, multiple lifetimes, each with its own general distribution, are considered to be alive simultaneously and the death of any one of them causes the process to move to another state. It is possible to model a wide variety of processes with a GSMP.

Matthes showed that a GSMP is insensitive with respect to a particular lifetime s (that is its stationary distribution depends on the distribution of the lifetime s only through its mean) if and only if a system of balance equations is satisfied by the stationary distribution when all lifetimes are taken to be exponential. Matthes' results, together with the work of König and Matthes [31] and König [29] are collected in König, Matthes and Nawrotzki [32].

A generalisation of Matthes' framework was given by König and Jansen [30] who introduced state-dependent speeds $c(s,g)$ at which the lifetime s is worked off when the state is g. The use of these speeds allowed functional dependencies between different lifetimes in the GSMP to be modelled. In particular, by putting a speed to zero it is possible to model the situation where a particular lifetime is not processed at all in some states.

Schassberger [38, 39, 40] proved Matthes' partial balance result in a different way utilising mixtures of Erlang distributions in place of the general distributions, extending these results to arbitrarily distributed lifetimes using weak continuity arguments. A general justification for the use of these arguments was given by Whitt [47]. Further results on insensitivity in GSMPs were given in Jansen, König and Nawrotzki [26], Burman [5], Franken, Arndt, König and Schmidt [13], Henderson [15] and Henderson and Taylor [16].

Whittle [49] provided a simple proof of the equivalence of partial balance and insensitivity, using a structure that initially appeared to be different to the GSMP. The simplicity of the proof derived in some part from the assumptions that Whittle made about his process. For example, Whittle's structure did not allow a general lifetime to immediately restart after having died, although it is fairly easy to modify the structure to include this feature. Whittle also implicitly assumed that all the speeds are positive, which does sustantially simplify the situation. It was trying to cope with these details that made the earlier proofs of König and Jansen [30] and Schassberger [41] somewhat more complicated than they otherwise would have been. When appropriate generalisations are included, Whittle's structure is completely equivalent to the GSMP, as was established by Schassberger [42] and Miyazawa [35]. Nonetheless it is notable for its elegance and simplicity. For this reason, it was used by the author in some of his own contributions to insensitivity theory [45, 46].

In a later paper, Whittle [50] combined his previous approach with the concept of weak coupling [48] to present an approach to insensitivity in terms of "imbedding". A Markov process Q with states n is said to be imbedded in a Markov process $\hat{Q}$ with a finer classification of states (n,x) if certain rules about the transitions of $\hat{Q}$ are obeyed and if $\pi(n) = \sum_x \pi(n,x)$ where $\pi(n)$ and $\pi(n,x)$ are the respective stationary distributions of Q and $\hat{Q}$. Using this concept, it follows that if a process $\hat{Q}$ is insensitive with respect to a set of general lifetimes then it imbeds a process with identical transition rates but negative exponential lifetimes. This follows by considering the finer classification x to be a set of supplementary variables describing the state of the general lifetimes. Whittle's results on insensitivity, imbedding and weak coupling were collected in his book [51].

The purpose of this paper is to provide an insight into the methods used in studying insensitivity and to act as a starting point for readers who are interested in learning more. We shall do this by looking in some detail at how insensitivity results are derived for the Erlang Loss System. This system has the advantage that it can be described as a symmetric queue as defined by Kelly [27, 28], or as a GSMP. It is thus an ideal vehicle to point out the similarities and the differences between the two approaches.

In Section 3.2, we shall start by looking at the Erlang Loss system as a symmetric queue. We shall follow this in Section 3.3 by looking at the same system within a GSMP framework. In Section 3.4 we shall illustrate how insensitivity results can be generalised to the queueing network context, and in Section 3.5 present a discussion of the very few non-standard insensitive systems that are known. A summary of our approach is given in Section 3.6.

3.2 The Erlang Loss System as a Symmetric Queue

Consider the M/G/C/C queue where the service time distribution G is allowed to be arbitrary. We shall assume that it has a density g on $[0,\infty)$, but our results can be shown to apply to the general case by the same weak continuity arguments [47] that were used to justify the extension from mixtures of Erlang distributions. Given that a service time has lasted for a time y, the probability that it finishes within a time interval δ is then given by $h(y)\delta + o(\delta)$ where the hazard function $h(y)$ is defined by

$$h(y) = \frac{g(y)}{1 - G(y)}. \tag{3.4}$$

We shall proceed by extending the definition of state so that the process still has a Markovian description. We can do this by labelling each of the individual customers and recording some information about the service at each one of them. Specifically, we shall record the spent service time of each of the customers that is present. Thus, instead of a continuous-time Markov chain with states n, we study a process with states $X(t)$ of the form $(n,y_1,\ldots,y_n)$ where y_i is the spent service time of the customer with label i. We shall stipulate that

Assumptions A

1. customers arriving when n customers are present are allocated each of the $n+1$ possible labels with probability $1/(n+1)$, and
2. when a customer departs, all customers with higher labels have their label decreased by one.

The resulting process is still Markovian, but it has a state space with continuous components. The infinitesimal generator

$$\lim_{t\to 0} \frac{d}{dt} E[f(X(t)) - f(X(0))],$$

which acts on a suitably-defined set of functions $\mathcal{F}$, is relatively easy to write down and we can proceed from this to a derivation of the stationary distribution. However, arguably, more insight is obtained by writing down equations that govern the probability densities $P(n,y_1,\ldots,y_n : t)$ that $X(t) = (n,y_1,\ldots,y_n)$. For a state $(n,y_1,\ldots,y_n)$ with $y_1,\ldots,y_n > 0$ and $0 < n < C$,

$$\begin{aligned}
&P(n,y_1+\delta,\ldots,y_n+\delta:t+\delta)\\
&=\left[1-[\lambda+\sum_{i=1}^{n}h(y_i)]\delta\right]P(n,y_1,\ldots,y_n:t)\\
&+\sum_{i=1}^{n+1}\int_0^\infty P(n+1,y_1,\ldots,y_i,z,y_{i+1},\ldots,y_n:t)h(z)\delta dz+o(\delta). \qquad (3.5)
\end{aligned}$$

The first term on the right hand side reflects the situation in which no arrival or departure occurs in the time interval $(t,t+\delta)$ and all that happens is that the spent service times of the customers who are present at time t age by an amount δ, while the summand in the second term reflects the situation in which there are $n+1$ customers present at time t and the one labelled $i+1$ departs, with all higher labels being decreased by one, as stipulated by Assumption A(2).

Dividing equation (3.5) by δ and letting $\delta\to 0$, we get

$$\begin{aligned}
&\frac{\partial P(n,y_1,\ldots,y_n:t)}{\partial t}+\sum_{i=1}^{n}\frac{\partial P(n,y_1,\ldots,y_n:t)}{\partial y_i}\\
&=-\left[\lambda+\sum_{i=1}^{n}h(y_i)\right]P(n,y_1,\ldots,y_n:t)\\
&+\sum_{i=1}^{n+1}\int_0^\infty P(n+1,y_1,\ldots,y_i,z,y_{i+1},\ldots,y_n:t)h(z)dz. \qquad (3.6)
\end{aligned}$$

When $n=C$, no arrivals can occur, nor can we ever be in state $n+1$, so the equations reduce to

$$\begin{aligned}
&\frac{\partial P(C,y_1,\ldots,y_C:t)}{\partial t}+\sum_{i=1}^{C}\frac{\partial P(C,y_1,\ldots,y_C:t)}{\partial y_i}\\
&=-\sum_{i=1}^{C}h(y_i)P(C,y_1,\ldots,y_C:t). \qquad (3.7)
\end{aligned}$$

When $n=0$, the equation is

$$\lambda P(0:t)=\int_0^\infty P(1,z:t)h(z)dz. \qquad (3.8)$$

To get the boundary conditions, we need to consider what happens at arrival instants, as stipulated by our Assumption A(1). For $0<n\le C$, we have

$$\begin{aligned}
&\int_0^\delta P(n,y_1+\delta,\ldots,y_i+\delta,u,\ldots y_n+\delta:t+\delta)du\\
&=\frac{\lambda\delta}{n}P(n-1,y_1,\ldots,y_n:t)+o(\delta).
\end{aligned}$$

Both sides of this equation describe the event that there were $n-1$ customers present at time t and a further customer arrived, and was allocated to position $i+1$, before time $t+\delta$. Again, dividing by δ and letting $\delta \to 0$, we get

$$P(n,y_1,\ldots,0,\ldots y_n : t) = \frac{\lambda}{n} P(n-1,y_1,\ldots,y_n : t). \tag{3.9}$$

To get equations for the stationary densities $\pi(n,y_1,\ldots,y_n)$, we put the time derivative to zero in equations (3.6), (3.7) and (3.8) which gives us

$$\sum_{i=1}^{n} \frac{\partial \pi(n,y_1,\ldots,y_n)}{\partial y_i} = -\left[\lambda + \sum_{i=1}^{n} h(y_i)\right] \pi(n,y_1,\ldots,y_n)$$
$$+ \sum_{i=1}^{n+1} \int_0^\infty \pi(n+1,y_1,\ldots,y_{i-1},z,y_i,\ldots,y_n) h(z)dz, \tag{3.10}$$

$$\sum_{i=1}^{C} \frac{\partial \pi(C,y_1,\ldots,y_C)}{\partial y_i} = -\sum_{i=1}^{C} h(y_i)\pi(C,y_1,\ldots,y_C) \tag{3.11}$$

and

$$\lambda \pi(0) = \int_0^\infty \pi(1,z)h(z)dz, \tag{3.12}$$

subject to

$$\pi(n,y_1,\ldots,0,\ldots y_n) = \frac{\lambda}{n} \pi(n-1,y_1,\ldots,y_n). \tag{3.13}$$

The proof of the insensitivity of the Erlang Loss model proceeds by verifying that a solution to the above set of equations is given by the product-form expression

$$\pi(n,y_1,\ldots,y_n) = \pi(0)\frac{\lambda^n}{n!} \prod_{i=1}^{n} (1-G(y_i)). \tag{3.14}$$

We do this by using the facts that, for any absolutely continuous G,

$$\frac{d(1-G(y))}{dy} = -g(y) = -h(y)(1-G(y)), \tag{3.15}$$

and so the ith term in the sum on the left hand side of both equations (3.10) and (3.11) is equal to the ith term in the sum on the right hand side. Also

$$\int_0^\infty (1-G(u))h(u)du = \int_0^\infty g(u)du = 1, \tag{3.16}$$

and so

$$-\lambda \pi(n,y_1,\ldots,y_n) + \sum_{i=1}^{n+1} \int_0^\infty \pi(n+1,y_1,\ldots,y_i,z,y_{i+1},\ldots,y_n)h(z)dz = 0. \tag{3.17}$$

This guarantees that the remaining terms of equation (3.10) and the two sides of equation (3.12) are equal. Also $G(0) = 0$, which ensures that (3.13) is satisfied.

A justification that the set of equations (3.10), (3.11) and (3.13) have a unique solution that sums to one, which indeed gives us the stationary densities of the Markov chain, follows from the work of Miyazawa and Yamazaki [36]. So (3.14) provides an expression for the stationary densities for the Markov process with the supplemented state space. It does depend on G, so where does the insensitivity come in? This occurs when we integrate out the supplementary variables. It is elementary that

$$\int_0^\infty (1-G(u))du = 1/\mu, \tag{3.18}$$

and so

$$\int_0^\infty \cdots \int_0^\infty \pi(n,y_1,\dots y_n)dy_n\dots dy_1 = \pi(0)\frac{\lambda^n}{\mu^n n!}, \tag{3.19}$$

which is identical to the probability (3.2) that there are n customers present in the system with exponential service times.

The decomposition of the stationary density of the supplemented process is typical of insensitive stochastic models. Consider the situation when lifetimes are exponential, but where we retain a notional labelling of customers. Then the flux $\lambda\pi(n-1)/n$ into state n due to the arrival of the customer with label i is equal to the flux $\mu\pi(n)$ out of state n due to the departure of the customer with label i. It is this relationship, which crucially depends on the assumption that customers arriving when n previous customers are present are allocated a label uniformly on the set $1,\dots,n+1$, that ensures that equation (3.13) is satisfied by the product-form expression (3.14). This relationship is reflected in the partial balance equations

$$\lambda\pi(n-1) = n\mu\pi(n), \tag{3.20}$$

which are a finer set of equations than the equations (3.1) that define the stationary distribution of the Markovian model that arises when service times are taken to be exponential. Insensitivity is usually associated with the satisfaction of some form of partial balance equation of this type.

3.3 The Erlang Loss System as a GSMP

In this section we shall discuss insensitivity in the Erlang Loss System by modelling it as a GSMP. We shall continue to denote the arrival rate by λ, the number of servers by C and assume that the service time distribution G has mean $1/\mu$ and a density g on $[0,\infty)$. The difference between the analysis here and that in Section 3.2 will lie in the method that we use for labelling lifetimes. Instead of shuffling labels up and down when customers arrive and depart in such a way that, when n customers are present, labels $1,\dots,n$ are all in use, we shall assign a label to each of the servers and denote the state by the subset of labels $\{1,\dots,C\}$ that are currently present.

So instead of denoting our supplemented states by $(n,y_1,\dots,y_n)$, reflecting the situation that there are n customers in the queue with the ith labelled customer having spent service time y_i, we shall use states of the form $(\phi,y_{s_1},\dots,y_{s_n})$, where $\phi = \{s_1,\dots,s_n\}$ is a subset of $\{1,\dots,C\}$ recording the labels of the servers at which customers are present and y_{s_i} is the spent service time of the customer at server s_i. It should be immediately apparent to the reader that this is a finer state classification than we used in Section 3.2: many states ϕ have n customers present. For $s \notin \phi$, we shall write $\phi + s$ for the state $\phi \cup \{s\}$, and for $s \in \phi$, we shall write $\phi - s$ for the state $\phi \setminus \{s\}$. Also we shall write $|\phi|$ for the number of elements in the set ϕ and ϕ_C for the set $\{1,\dots,C\}$.

As in Section 3.2, we need to make an assumption about how arriving customers are allocated to servers. Thus we have

Assumption B

1. Customers arriving when the state is ϕ are allocated to each of the $C - |\phi|$ free servers with probability $1/(C - |\phi|)$.

No assumption about what happens when a customer departs is necessary: the label of the corresponding server is simply deleted from the current state.

When the service times are exponential, it is unnecessary to keep track of the spent service times and the stationary probabilities of the resulting Markov chain satisfy the equations

$$\lambda\pi(\emptyset) = \sum_{i=1}^{C} \mu\pi(\{i\}) \tag{3.21}$$

$$(\lambda + |\phi|\mu)\pi(\phi) = \sum_{s\in\phi} \frac{\lambda}{C - |\phi| + 1}\pi(\phi - s) + \sum_{s\notin\phi} \mu\pi(\phi + s) \tag{3.22}$$

$$C\mu\pi(\phi_C) = \sum_{s=1}^{C} \lambda\pi(\phi_C - s). \tag{3.23}$$

The solution that sums to unity is

$$\pi(\phi) = \pi(0)\frac{\lambda^{|\phi|}(C - |\phi|)!}{\mu^{|\phi|}C!} \tag{3.24}$$

where

$$\pi(0) = \frac{1}{\sum_{i=0}^{C} \rho^i i!} \tag{3.25}$$

and, as in Section 3.1, $\rho = \lambda/\mu$.

For any ϕ, the expression on the right hand side of equation (3.24) depends on ϕ only through $|\phi|$ and so, conditional on the fact that $|\phi| = n$ the distribution is uniform. Summing over the $\binom{C}{n}$ states ϕ with $|\phi| = n$, we see that the probability that there are n customers present in the system is

$$\pi(n) = \pi(0)\frac{\lambda^n}{\mu^n n!}, \tag{3.26}$$

which agrees with equation (3.2).

When the service times are generally-distributed, we can supplement the state by the spent service time of the customers at each of the servers and use an approach similar to that of Section 3.2 to derive equations for the stationary densities $\pi(\phi, y_{s_1}, \ldots, y_{s_n})$ of the supplemented system. These are

$$\lambda\pi(\emptyset) = \sum_{i=1}^{C}\int_0^\infty \pi(\{i\}, y_i)h(y_i)dy_i \tag{3.27}$$

$$\sum_{s\in\phi}\frac{\partial}{\partial y_s}\pi(\phi, y_{s_1}, \ldots, y_{s_n}) = -\left[\lambda + \sum_{s\in\phi} h(y_{s_i})\right]\pi(\phi, y_{s_1}, \ldots, y_{s_n}) \tag{3.28}$$

$$+\sum_{s\notin\phi}\int_0^\infty \pi(\phi+s, y_{s_1}, \ldots, y_{s_n}, y_s)h(y_s)dy_s$$

$$\pi(\phi+s, y_{s_1}, \ldots, y_{s_n}, 0) = \frac{\lambda}{C-n}\pi(\phi, y_{s_1}, \ldots, y_{s_n}) \tag{3.29}$$

$$\sum_{s\in\phi_C}\frac{\partial}{\partial y_s}\pi(\phi, y_{s_1}, \ldots, y_{s_n}) = -\sum_{s\in\phi} h(y_{s_i})\pi(\phi_C, y_{s_1}, \ldots, y_{s_n}) \tag{3.30}$$

where $n = |\phi|$. Like equations (3.10), (3.11) and (3.13), equations (3.27), (3.28) and (3.29) and (3.30) have a solution that factorises into a product form over the generally distributed lifetimes. This is

$$\pi(\phi, y_{s_1}, \ldots, y_{s_n}) = \pi(0)\frac{\lambda^n(C-n)!}{C!}\prod_{i=1}^{n}(1 - G(y_{s_i})). \tag{3.31}$$

Integrating out the supplementary variables gives us,

$$\int_0^\infty \cdots \int_0^\infty \pi(\phi, y_{s_1}, \ldots, y_{s_n})dy_{s_n}\ldots dy_{s_1} = \pi(0)\frac{\lambda^n(C-n)!}{\mu^n C!}, \tag{3.32}$$

which demonstrates the insensitivity. The fact that (3.31) satisfies equations (3.27), (3.28) and (3.29) and (3.30) again follows from a partial balance result. Specifically, the stationary distribution of the Markovian system when all service times are taken to be exponential not only satisfies equations (3.21), (3.22) and (3.23) but it also satisfies the finer balance equations

$$\frac{\lambda}{C-|\phi|}\pi(\phi) = \mu\pi(\phi+s) \tag{3.33}$$

for all $\phi \neq \phi_C$. This embodies that notion that, in the Markovian system, the probability flux into state $\phi+s$ due to the arrival of the customer at server s is equal to the probability flux out of state $\phi+s$ due to the departure of the customer at server s.

There are some subtle differences between the insensitivity result that we have discussed in this section and the one that we proved in Section 3.2. First, observe that expression (3.32) implies insensitivity of the stationary probability that the state of the queue is ϕ. This is a stronger statement than saying that the stationary probability that there are n customers present is insensitive.

Second, under the formulation of Section 3.2, particular lifetimes change their label when arrivals and departures occur whereas, in the formulation of this section, lifetimes retain the label for their entire duration. A consequence of this is that, in the model of Section 3.2, all service times have to be chosen from the same distribution while, in the formulation of this section, service times at a particular server can have their own server-specific distribution. These distributions need not even have the same mean, although this would necessitate a change to the form of the stationary distribution (3.24). While this will not usually present extra flexibility from a modelling point of view, since we usually want customers at a queue to select their service times from the same distribution, this distinction illustrates that the result in this section can be thought of as being more general that that of Section 3.2.

This observation might lead us to ask why we do not always use a GSMP formulation rather than a symmetric queue formulation. The reason for this is that the GSMP formulation works only when there are finitely-many possible labels available to be distributed to the lifetimes. This is the case for queueing models when there is finite waiting room, but not when the possible number in the queue is unbounded. In this case, it is impossible for an arriving customer to choose a label (that is a server) uniformly from infinitely-many possibilities, as would be required by the analogue of Assumption B. Without this assumption, it is not possible to show that equations analogous to (3.27) to (3.30) are satisfied by the product-form distribution (3.31). The only alternative is the relabelling up and down of customers at arrival and departure points that we used in our formulation of Section 3.2. In this way, labels need only be considered for customers that are actually present in the queue, rather than for all the customers that could potentially come to the queue. This issue was discussed in more detail by Barbour [2] and Schassberger [42].

3.4 Insensitive Queueing Networks

As we mentioned in Section 3.1, one of the major reasons that the study of insensitive systems became popular in the last part of the twentieth century was that some classes of queueing network with wide applicability were shown to be insensitive. In particular, the networks studied by Baskett, Chandy, Muntz and Palacios [4] and Kelly [27, 28] were used as models for a number of different systems of engineering significance, particular in telecommunications and computer networking.

The essential observation of [4, 27, 28] and the other papers on insensitive queueing networks that followed them (see Section 3.1 for a discussion) is that many types of insensitive queue can be inserted into a network and the whole system will retain the insensitivity property. Because it has finite capacity, the Erlang loss queue is not a queue that can be inserted into such a network, at least in a natural way, so we shall illustrate this phenomenon using the $M/G/\infty$ queue, which we can think of as a 'loss queue with infinite capacity'. The model that we shall discuss below is simpler that that used in the work of either Baskett, Chandy, Muntz and Palacios [4] or Kelly [27, 28], who allowed for queues to contain multiple customer types and for the routing to depend on customer type. However, our analysis contains the essential idea behind the justification for insensitivity in all the queueing network models in which it is known to occur, and the context is simple enough for the reasoning to be straightforward.

Consider an $M/G/\infty$ queue at which customers arrive according to a Poisson process with parameter λ, and where the service time distribution G has density g and mean $1/\mu$. Assuming that labels are allocated to customers in accord with assumption A, an analysis similar to that in Section 3.2 can be used show that the stationary probability density $\pi(n,y_1,\ldots,y_n)$ that there are n customers present, and the spent service time of the customer with label i is y_i, is given by

$$\pi(n,y_1,\ldots,y_n) = \pi(0)\frac{\lambda^n}{n!}\prod_{i=1}^{n}(1-G(y_i)). \tag{3.34}$$

where $\pi(0)$ is equal to $\exp(-\rho)$. By integrating with respect to the y_i, we can see that the stationary probability density that there are n customers present depends on G only through the fact that its mean is $1/\mu$, and so this queue is insensitive.

Now consider a finite collection of K such queues, the kth of which has external arrivals following a Poisson process with parameter λ_k, and where the service time distribution G_k has density g_k, hazard function h_k and mean $1/\mu_k$. When a customer completes service at queue j, it moves to queue k with probability r_{jk} or departs the network entirely with probability r_{j0}, where $\sum_{k=0}^{K} r_{jk} = 1$. These routing probabilities are such that every customer eventually will, with probability one, depart the network.

When the service times are exponentially-distributed the whole system can be modelled by a continuous-time Markov chain with states $n = (n_1,\ldots,n_K)$, with n_k the number of customers in queue k, and stationary distribution $\pi(n)$ that satisfies the equations

$$\begin{aligned}\sum_{k=1}^{K}(\lambda_k + n_k\mu_k)\pi(n) &= \sum_{k=1}^{K}\lambda_k I(n_k>0)\pi(n-e_k) + \sum_{k=1}^{K}(n_k+1)\mu_k\pi(n+e_k)r_{k0}\\ &\quad+\sum_{k=1}^{K}\sum_{j\neq k}(n_j+1)\mu_j I(n_k>0)\pi(n+e_j-e_k)r_{kj},\end{aligned} \tag{3.35}$$

where e_k is the unit vector with a one in the kth position. Equations (3.35) have solution

$$\pi(n) = \prod_{k=1}^{K}\exp(-\eta_k)\frac{\eta_k^{n_k}}{n_k!}, \tag{3.36}$$

where $\eta_k = \alpha_k/\mu_k$ and $\alpha = (\alpha_1, \ldots, \alpha_K)$ satisfies the traffic equations

$$\alpha_k = \lambda_k + \sum_{j \neq k} \alpha_j r_{jk}. \tag{3.37}$$

Summation of equation (3.37) over k gives us the result that

$$\sum_{k=1}^{K} \lambda_k = \sum_{k=1}^{K} \alpha_k r_{k0}. \tag{3.38}$$

The key to establishing the product-form stationary distribution (3.36) is to use equations (3.37) and (3.38) to show that, for all states n and queues k, this expression satisfies the partial balance equations

$$n_k\mu_k\pi(n) = \lambda_k I(n_k > 0)\pi(n - e_k) + \sum_{j \neq k} (n_j + 1)\mu_j I(n_k > 0)\pi(n + e_j - e_k) r_{jk}, \tag{3.39}$$

and

$$\sum_{k=1}^{K} \lambda_k \pi(n) = \sum_{k=1}^{K} (n_k + 1)\mu_k \pi(n + e_k) r_{k0}. \tag{3.40}$$

This is essentially the result of Jackson [20] that the stationary distribution of a network of $M/M/\infty$ queues factorises into a product over the queues. The fact that the stationary distribution, defined to be the solution to equations (3.35), also satisfies the partial balance equations (3.39) and (3.40) has important implications for insensitivity. To see this, we need to expand the state description of the network so that each service time has a label of its own.

Let the state of the network of $M/G/\infty$ queues be defined by $(n, y_1, \ldots, y_K)$ where the vector $y_k = (y_{k1}, \ldots, y_{kn_k})$ contains the spent service times of the customers at queue k. We assume that labels are allocated to customers at queue k in accord with assumption A, whether the arriving customer comes from outside the network or from another queue. Using reasoning similar to that in Section 3.2, we can show that the stationary density of the queueing network satisfies the partial differential equations

$$\begin{aligned} \sum_{k=1}^{K} \sum_{i=1}^{n_k} \frac{\partial \pi(n, y_1, \ldots, y_K)}{\partial y_{ki}} &= -\sum_{k=1}^{K} \left[\lambda_k + \sum_{i=1}^{n_k} h_k(y_{ki}) \right] \pi(n, y_1, \ldots, y_K) \\ &+ \sum_{k=1}^{K} \sum_{i=1}^{n_k+1} \int_0^\infty \pi(n + e_k, y_1, \ldots, y_k + z_i, \ldots y_K) h_k(z) r_{k0} dz, \end{aligned} \tag{3.41}$$

where $y_k + z_i$ is short-hand notation for the vector $(y_{k1}, \ldots, y_{k(i-1)}, z, y_{ki}, \ldots, y_{kn_k})$. The boundary conditions are

$$\pi(n+e_k,y_1,\ldots,y_k+0_i,\ldots,y_K) = \frac{\lambda_k}{n_k+1}\pi(n,y_1,\ldots,y_k,\ldots,y_K)$$
$$+\sum_{j\neq k}\frac{1}{n_k+1}\sum_{i=1}^{n_j+1}\int_0^\infty \pi(n+e_j,y_1,\ldots,y_j+z_i,\ldots y_K)h_j(z)r_{jk}dz, \qquad (3.42)$$

where the n_k+1 in the denominator arises because a customer arriving at queue k will choose any one of the n_k+1 available labels with equal probability. The solution to equations (3.41) and (3.42) is

$$\pi(n,y_1,\ldots,y_K) = \prod_{k=1}^{K}\left[\exp(-\eta_k)\frac{\alpha_k^{n_k}}{n_k!}\prod_{i=1}^{n_k}(1-G_k(y_{ki}))\right]. \qquad (3.43)$$

This can be established by observing that relations analogous to equations (3.15) and (3.16) hold at each of the queues. The relation corresponding to (3.15) implies that the (k,i)th partial derivative on the left hand side of equation (3.41) is balanced by the (k,i)th term of the form $h_k(y_{ki})\pi(n,y_1,\ldots,y_K)$ on the right hand side. The relation corresponding to (3.16), together with (3.38), implies that $-\sum_{k=1}^K \lambda_k\pi(n,y_1,\ldots,y_K)$ on the right hand side of equation (3.41) is balanced by the integral term $\sum_{k=1}^K\sum_{i=1}^{n_k+1}\int_0^\infty \pi(n+e_k,y_1,\ldots,y_k+z_i,\ldots y_K)h_k(z)r_{k0}dz$. Using equation (3.37), it can easily be verified that expression (3.43) satisfies equations (3.42).

Integration of the stationary distribution (3.43) with respect to all of the spent lifetimes y_{ki} gives us the fact that the stationary distribution of the occupancies at each of the nodes is given by (3.36) for all choices of service time distributions $\{G_k\}$ that have means $\{\mu_k\}$. It is instructive to think about the factors that lead to this insensitivity result. These are the facts that

- that the stationary distribution of the Markovian network factorises into a product form over the queues, and
- that the stationary distribution of the individual queues, with state spaces including information on the spent service time of each customer, factorise into product forms over the individual customers.

These factorisation properties have, in turn, arisen from partial balance properties:

- that the transition flux into state n of the Markovian network due to an arrival at queue k is balanced by the transition flux out of state n due to a departure at queue k, and
- that the transition flux into a state n_k due to the arrival of the customer with label i is balanced by the transition flux out of state n_k due to the departure of the customer with label i.

We can see that the two partial balance properties have led to the fact that the network is insensitive. The second requirement can be shown to be satisfied by any symmetric queue and thus any such queue can serve as a component in an insensitive network. The first requirement follows from the fact that the routing of customers is ‘of Jackson type’ and, in particular, that the rate of transition of customers from queue j to queue k does not depend on the state at queue k. It is possible for these types of partial balance to be traded off in a restricted way and for a network still to be insensitive. For example, if the routing matrix $R = [r_{jk}]$ is the transition matrix of a reversible discrete-time Markov chain, then some types of dependence on the state of the destination node, including blocking, can be incorporated (see, for example, [17]). However, in general, the statement that insensitivity in a queueing network is associated with partial balance holds. In the next section, we shall discuss the few known examples of insensitivity that are not associated with partial balance.

3.5 Non-Standard Insensitive Models

In some queueing systems there has been observed a form of insensitivity which does not fall into the class of insensitivity discussed above. This type of insensitivity is not associated either with a product form of the supplemented stationary distribution or with partial balance. The first example of such insensitivity was discussed by Jacobi [21], who showed that an Erlang loss system with one overflow server is insensitive.

Another example was given by Wolff and Wrightson [52] who generalised a system which was considered earlier by Chaiken and Ignall [6]. Wolff and Wrightson's system has two arrival streams to a two server loss system with stream-dependent service time distribution. Stream 1 has preference for server 1, while stream 2 has preference for server 2. If the state of the system were defined as the number of customers of each type in the system then it would be an ordinary Erlang loss system with two types of customer, which can be shown to be insensitive and to possess product form using techniques similar to those discussed in Sections 3.2 and 3.3.

Wolff and Wrightson showed, however, that this system is still insensitive if the states are defined by the busy servers, irrespective of which type of customer is present at the server. It is interesting to note that the system with either of these state definitions can be obtained by amalgamating states of a refined GSMP in which both the position and type of each customer are recorded in the state space. The stationary distribution of this refined GSMP does not satisfy the partial balance equations analogous to equation (3.33) and hence is not insensitive. It appears that the distribution-dependent components of the stationary state probabilities of the refined process cancel out if the states are amalgamated according to either type or position.

Jacobi's [21] result was built upon by Jansen [23] to come up with a class of queues which are insensitive but do not possess product form over the supplementary variables. Jansen considered an Erlang loss system with C servers, m Poisson arrival streams, with stream i having rate λ_i, and states $\phi \subseteq \{1,\ldots,C\}$ defining the busy servers. He defined $q_j(\phi,i)$ as the probability that a type-i customer, arriving to find a state ϕ, is allocated to server j, and derived a set of conditions on the $q_j(\phi,i)$, sufficient for the process to be insensitive. Jansen's conditions were slightly incorrect. However it is reasonably easy to show that the correct set of conditions is: For $j \in \{1,\ldots,C\}$ and $i \in \{1,\ldots,m\}$,

$$q_j(\phi,i) \leq \frac{1}{C-1} \tag{3.44}$$

for all ϕ such that $|\phi| < C-1$, for $j \notin \phi$ and for all i,

$$q_j(\phi,i) = \frac{1}{C-|\phi|}\sum_{k\in\phi} p_k + \frac{1}{C-|\phi|-1}\sum_{k\notin\phi} p_k \tag{3.45}$$

where

$$p_k = \frac{\sum_{l=1}^m \lambda_l\,[1-(C-1)q_k(\phi,l)]}{\sum_{l=1}^m \lambda_l} \tag{3.46}$$

and, for all i and j

$$q_j(\{1,\ldots,C\}-\{j\},i) = 1. \tag{3.47}$$

Jansen's class of systems includes loss queues where the allocation of customers to servers is "nearly random". In particular if we take $C = 2$ so that $q_j(\phi,i) \leq 1$ we have complete freedom to choose allocation probabilities. A choice of $q_j(\phi,i) = \delta_{ij}$ gives Wolff and Wrightson's system.

The common thread in all of these "non-standard" insensitive systems is that they apply in models where certain lifetimes are constrained to have the same distribution. Although partial balance is necessary and sufficient for the standard type of insensitivity, the existence of these systems shows that we can have insensitivity without partial balance if we insist that certain lifetimes have common distributions. The only (to the author's knowledge) clue to the form of the stationary distribution for such a system, supplemented by variables to describe spent or residual lifetimes, was given by Henderson [14], who presented a solution for the Laplace transform of the equations for the supplemented stationary distribution of Wolff and Wrightson's model. Unfortunately, Miyazawa and Yamazaki [36] pointed out that Henderson omitted the necessary step of verifying that the inverse Laplace transform of his solution can be interpreted as a distribution function. Furthermore they presented a family of solutions similar to Henderson's. Since it is not possible for every member of this family to be the Laplace transform of the supplemented stationary distribution of Wolff and Wrightson's model, it is clear that the form of this distribution is yet to be resolved.

3.6 Conclusion

In this paper, we have presented an introduction to insensitivity as it occurs in stochastic models. Our approach has been to illustrate the main ideas using simple special cases. Thus, in Sections 3.2 and 3.3, we illustrated insensitivity in symmetric queues and insensitivity in GSMPs both within the context of an Erlang Loss model. In Section 3.4 we used a network of infinite server queues to illustrate insensitivity in a product-form queueing network. Finally, in Section 3.5 we discussed non-standard insensitive models in which insensitivity is not associated either with partial balance or a product-form supplemented stationary distribution.

Acknowledgement

The author gratefully acknowledges the support of the Australian Research Council Centre of Excellence for the Mathematics and Statistics of Complex Systems (MASCOS).

References

1. Barbour A., Networks of Queues and the Method of Stages, *Advances in Applied Probability*, **8** (1976), 584–591.
2. Barbour A., Generalised Semi-Markov Schemes and Open Queueing Networks, *Journal of Applied Probability*, **19** (1982), 469–474.
3. Barbour A. and Schassberger R., Insensitive Average Residence Times in Generalised Semi-Markov Processes, *Advances in Applied Probability*, **13** (1981), 720–735.
4. Baskett F., Chandy K., Muntz R. and Palacios J., Open, Closed and Mixed Networks of Queues with Different Classes of Customers, *Journal of the Association for Computing Machinery*, **22** (1975), 248–260.
5. Burman D., Insensitivity in Queueing Systems, *Advances in Applied Probability*, **13** (1981), 846–859.
6. Chaiken J. and Ignall E., An Extension of Erlang's Formulas which Distinguishes Individual Servers, *Journal of Applied Probability*, **9** (1972), 192–197.
7. Chandy K., Howard J. and Towsley D., Product Form and Local Balance in Queueing Networks, *Journal of the Association for Computing Machinery*, **24** (1977), 250–263.
8. Chandy K. and Martin J., A Characterisation of Product Form Queuing Networks, *Journal of the Association for Computing Machinery*, **30** (1983), 286–299.
9. Cohen J. (1957), The Generalised Engset Formulae, *Philips Telecommunication Revue*, **18** (1957), 158–170.
10. Erlang A., Solution of some Problems in the Theory of Probabilities of Significance in Automatic Telephone Exchanges, *Post Office Electrical Engineer's Journal,* **10** (1917), 189–197.
11. Fakinos D., The M/G/k Group-Arrival Group-Departure Loss System, *Journal of Applied Probability*, **19** (1982), 826–834.
12. Fortet R., *Calcul des probabilités*, Centre National de la Recherche Scientifique, Paris (1950).
13. Franken P., Arndt U., König D. and Schmidt V., *Queues and Point Processes*, Wiley, London (1982).
14. Henderson W., Non Standard Insensitivity, *Journal of Applied Probability*, **20** (1983), 288–296.
15. Henderson W., Insensitivity and Reversed Markov Processes, *Advances in Applied Probability*, **15** (1983), 752-768.
16. Henderson W. and Taylor P.G., Insensitivity of Processes with Interruptions, *Journal of Applied Probability*, **26** (1989), 242-258.
17. Hordijk A. and Van Dijk N.M., Networks of Queues with Blocking, *Performance 81*, (ed K. Kylstra), North Holland (1981), 51–65.
18. Hordijk A. and Van Dijk N.M., Networks of Queues, *Proceedings of the International Seminar on Modelling and Performance Evaluation Methodology*, INRIA, **1** (1983), 79–135.

19. Hordijk A. and Van Dijk N.M., Adjoint Processes, Job Local Balance and Insensitivity for Stochastic Networks, *Bulletin of the 44th Session International Statistical Institute*, **50** (1983), 776–788.
20. Jackson J. Networks of Waiting Lines, *Operations Research*, **5** (1957) 518–521.
21. Jacobi H., Eine Unempfindlichkeitseigenschaft für geordnete Bündel ungeordnete Teilbündel, *Wissenschaft Zeitschrift Friedrich-Schiller-Universität Jena, Mathematische-Naturische*, **14** (1965), 251–260.
22. Jagerman D.L., Some Properties of the Erlang Loss Function, *Bell System Technical Journal*, **53** (1974), 525–557.
23. Jansen U., Unempfindlichkeitseigenschaft für verschiedene Auswahlregeln ohne vorliegen der Produktform der stationären Verteilung, *Elektronische Informationsverarbeitung Kybernetik*, **16** (1980), 443–448.
24. Jansen U., Conditional Expected Sojourn Times in Insensitive Queueing Systems and Networks, *Advances in Applied Probability*, **15** (1984), 752–768.
25. Jansen U. and König D., Insensitivity and Steady-State Probabilities in Product Form for Queueing Networks, *Elektronische Informationsverarbeitung Kybernetik*, **16** (1980), 385–397.
26. Jansen U., König D. and Nawrotzki K., A Criterion of Insensitivity for a Class of Queueing Systems with Random Marked Point Processes, *Mathematische Operationsforschung und Statistik. Series Optimization*, **10** (1979), 379–403.
27. Kelly F., Networks of Queues, *Advances in Applied Probability*, **8** (1976), 416–432.
28. Kelly F., *Reversibility and Stochastic Networks*, Wiley, London (1979).
29. König D., Verallgemeinerungen der Engsetschen Formeln, Mathematische Nachrichten, **28** (1965), 145–155.
30. König D. and Jansen U., Stochastic Processes and Properties of Invariance for Queueing Systems with Speeds and Temporary Interruptions, *Transactions of the 7th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, (1974), 335–343.
31. König D. and Matthes K., Verallgemeinerungen der Erlangschen Formeln I, Mathematische Nachrichten, **26** (1963), 45–56.
32. König D., Matthes K. and Nawrotzki K., *Verallgemeinerungen der Erlangschen und Engsetschen Formeln (Eine Methode in der Bedienungstheorie)*, Akademie-Verlag, Berlin (1967).
33. Kosten L., On the Validity of the Erlang and Engset Loss Formulae, Het P.T.T. Bedrijf, **2** (1942), 42–45.
34. Matthes K., Zur Theorie der Bedienungsprozesse, *Transactions of the 3rd Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, (1962), 513–528.
35. Miyazawa M., The Characterization of the Stationary Distribution of the Supplemented Self-Clocking Jump Process, *Mathematics of Operations Research*, **16** (1991), 547-565.
36. Miyazawa M. and Yamazaki G., The Basic Equations for a Supplemented GSMP and its Applications to Queues, *Journal of Applied Probability*, **25** (1988), 565–578.
37. Noetzel A., A Generalized Queueing Discipline for Product Form Queueing Networks, *Journal of the Association for Computing Machinery*, **26** (1979), 779–793.
38. Schassberger R., On the Equilibrium Distribution of a Class of Finite-State Generalized Semi-Markov Processes, *Mathematics of Operations Research*, **1** (1976), 395–406.
39. Schassberger R., Insensitivity of Steady-State Distributions of Generalized Semi-Markov Processes. Part 1., *Annals of Probability*, **5** (1977), 87–89.
40. Schassberger R., Insensitivity of Steady-State Distributions of Generalized Semi-Markov Processes. Part II., *Annals of Probability*, **6** (1978), 85–93.
41. Schassberger R., Insensitivity of Steady-State Distributions of Generalised Semi-Markov Processes with Speeds, *Advances in Applied Probability* **10** (1978), 836–851.
42. Schassberger R., Two Remarks on Insensitive Stochastic Models, *Reihe Mathematik*, Technische Universität Berlin (1985).

43. Sevastyanov B., An Ergodic Theorem for Markov Processes and its Application to Telephone Systems with Refusals, *Theory of Probability and its Applications*, **2** (1957), 104–112.
44. Takacs L., On Erlang's Formula, *Annals of Mathematical Statistics*, **40** (1969), 71–78.
45. Taylor P.G., *Aspects of Insensitivity in Stochastic Processes*, PhD Thesis, University of Adelaide, 1987.
46. Taylor P.G., Insensitivity in Processes with Zero Speeds, *Advances in Applied Probability*, **21** (1989), 612–628.
47. Whitt W., Continuity of Generalised Semi-Markov Processes, *Mathematics of Operations Research*, **5** (1980), 494–501.
48. Whittle P., Weak Coupling in Stochastic Systems, *Proceedings of the Royal Society, Series A*, **395** (1984), 141–151.
49. Whittle P., Partial Balance and Insensitivity, *Journal of Applied Probability*, **22** (1985), 168–176.
50. Whittle P., Partial Balance, Insensitivity and Weak Coupling, *Advances in Applied Probability*, **18** (1986) , 706-723.
51. Whittle P., *Systems in Stochastic Equilibrium*, Wiley, London (1986).
52. Wolff R. and Wrightson C., An Extension of Erlang's Loss Formula, *Journal of Applied Probability*, **13** (1976), 628–632.

Chapter 4
Palm Calculus, Reallocatable GSMP and Insensitivity Structure

Masakiyo Miyazawa

Abstract This chapter discusses Palm calculus and its applications to various processes including queues and their networks. We aim to explain basic ideas behind this calculus. Since it differs from the classical approach using Markov processes, we scratch from very fundamental facts. The main target of Palm calculus is stationary processes, but we are also interested in its applications to Markov processes. For this, we consider piece-wise deterministic processes and reallocatable generalized Markov processes, RGSMP for short, and characterize their stationary distributions using Palm calculus. In particular, the insensitive structure of RGSMP with respect to the lifetime distributions of its clocks is detailed. Those results are applied to study the insensitive structure of product form queueing networks with respect to service requirement distributions.

4.1 Introduction

In queues and their networks, it is typical that their time evolutions substantially change only when customers arrive or complete service. That is, essential changes occur only at embedded instants on the continuous time axis. This is a prominent feature of stochastic models for those systems. Such time instants are caused typically by arrivals and departures of customers, and often called discrete events. As is well known, it motivates to use discrete time stochastic processes embedded at those time instants. However, sample paths of such embedded processes may lose key information on the system evolution. Thus, they may be only useful in limited situations. Of course, those sample paths can retain full information of the system if we supplement them with all information between the embedded instants. However,

Masakiyo Miyazawa
Tokyo University of Science
e-mail: miyazawa@is.noda.tus.ac.jp

R.J. Boucherie and N.M. van Dijk (eds.), *Queueing Networks: A Fundamental Approach*,
International Series in Operations Research & Management Science 154,
DOI 10.1007/978-1-4419-6472-4_4, © Springer Science+Business Media, LLC 2011

it causes their descriptions to be complicated, and therefore analytical tractability may be lost.

In this chapter, we introduce a stochastic model to capture those discrete time nature in the continuous time setting. We aim to avoid to simultaneously use different processes for describing discrete events under the continuous time setting. Instead of doing so, we introduce different probability measures on the same sample space. They describe observations at times of interests. That is, they are used to compute characteristics of system states at continuous time or various embedded instants. When such characteristics are expectations of some random quantities, they are called time or event averages, respectively. Under certain stationary assumptions, it is shown that those probability measures are nicely related. This leads to useful relationship among time and event averages. It is referred to as Palm calculus since the probability measures concerning embedded epochs are called Palm distributions.

We apply this Palm calculus to stochastic processes arisen in queues and their networks. However, the Palm calculus itself may not be convenient since it usually involves integrations over the time axis. To ease this, we introduce a rate conservation law, which may be considered as a differential form of the Palm calculus.

Those results by the Palm calculus are very general in the sense that they only require the stationary assumption. However, we may need more specific models to compute characteristics in closed form. For this, we consider a piece-wise deterministic Markov process, PDMP for short. We then specialize it as a reallocatable generalized semi-Markov process, RGSMP for short. We are interested in when RGSMP has a certain nice form of the stationary distribution. It turns out that conditions for this form are closely related to those for a queueing network to have a product form stationary distribution. Under the same conditions, we also consider the conditional mean sojourn time of a customer in a queue or in a network given total amount of his/her work.

In this chapter, we consider analytical tools for studying queueing models rather than just to collect results for applications. However, we divide this chapter into small sections to highlight each topic. We expect the reader has some background in introductory levels of the probability and measure theories. Some descriptions particularly in the first few sections may look too formal since they are different from those in the standard queueing literature. However, arguments are essentially elementary. A problem is probably in the language that we use. So, we provide its details.

4.2 Shift operator group

When we consider a stochastic process for queueing models, we usually do not explain the probability space, i.e., the triplet of a sample space, a σ-field and a probability measure on it under which the process is defined. This is because the probability space is obviously identified. However, we here start from this very basic description since we will consider different probability measures on the same sample space and σ-field. We also play with different stochastic processes which have a common time axis. For this, it is convenient to implement a time axis in the sample space. Thus, we introduce a time-shift operator on it.

Let $(\Omega, \mathcal{F})$ be a measurable space, and let θ_t be an operator on Ω, i.e., a mapping from Ω to Ω for each real number $t \in \mathbb{R}$. Since we consider a function of time t and analytic operations on it, we need conditions to well define them. Thus, we formally define the operator θ_t in the following way.

Definition 4.1. For the operator θ_t on Ω for each t, define function φ from $\mathbb{R} \times \Omega$ to Ω by $\varphi(t, \omega) = \theta_t(\omega)$, and let $\mathcal{B}(\mathbb{R})$ be the Borel σ-field on $\mathbb{R}$. If the condition:

(4.2a) φ is $\mathcal{B} \times \mathcal{F}/\mathcal{F}$-measurable, i.e., $\varphi^{-1}(A) \in \mathcal{B} \times \mathcal{F}$ for all $A \in \mathcal{F}$,

is satisfied, then $\{\theta_t; t \in \mathbb{R}\}$ is said to be measurable. In addition to this, if

(4.2b) For any $s,t \in \mathbb{R}$, $\theta_s \circ \theta_t = \theta_{s+t}$, namely,

$$\theta_s(\theta_t(\omega)) = \theta_{s+t}(\omega), \qquad \omega \in \Omega,$$

is satisfied, then $\{\theta_t\}$ is said to be a shift operator group. □

Example 4.1. A natural candidate for the sample space Ω for θ_t to be defined is the set of functions on $\mathbb{R}$, which represents the time axis. For example, let S be a complete, separable metric space, which is called Polish space, and let $\mathcal{B}(S)$ be the Borel σ-field on S. Thus, we have measurable space $(S, \mathcal{B}(S))$. Let Ω be the set of S-valued functions on $\mathbb{R}$ whose discontinuous points are countable at most and are right continuous. Since $\omega \in \Omega$ is a function of time, it can be written as $\{\omega(t)\}$. The σ-field $\mathcal{F}$ can be generated from all the following subsets of Ω for all $n \geq 1, t_i \in \mathbb{R}, B_i \in \mathcal{B}(S)$ for $i = 1, 2, \ldots, n$.

$$\{\omega \in \Omega; \omega(t_i) \in B_i, i = 1, 2, \ldots, n\}.$$

Then, we can define the shift operator group θ_t through

$$\theta_t(\omega)(s) = \omega(s+t), \qquad s,t \in \mathbb{R}.$$

We refer to this θ_t as a natural shift operator. □

We next define stationarity with respect to the shift operator.

Definition 4.2. Let $\{\theta_t\}$ be an operator group on a measurable space $(\Omega, \mathcal{F})$. If a probability measure on $(\Omega, \mathcal{F})$ satisfies

$$P(\theta_t^{-1}(A)) = P(A), \qquad t \in \mathbb{R}, A \in \mathcal{F},$$

then P is said to be θ_t-stationary, or stationary with respect to $\{\theta_t\}$. □

Up to now, we have only considered the time to be real valued, i.e., continuous. We are also interested in discrete time. In this case, the shift operator on Ω is denoted by η_n for $n \in \mathbb{Z}$, where $\mathbb{Z}$ is the set of all integers. Obviously, conditions (4.2a) and (4.2b) are replaced by

(4.2c) For any $n \in \mathbb{Z}$, $\eta_n^{-1}(A) \in \mathcal{F}$ for $A \in \mathcal{F}$,
(4.2d) For any $m,n \in \mathbb{Z}$, $\eta_m \circ \eta_n = \eta_{m+n}$.

Similarly to θ_t, $\{\eta_n; n \in \mathbb{Z}\}$ is said to be measurable if (4.2c) is satisfied, and said to be an discrete time shift operator group if (4.2c) and (4.2d) are satisfied. Furthermore, P is said to be η_n-stationary if $P(\eta_1^{-1}(A)) = P(A)$ for all $A \in \mathcal{F}$.

We next apply the shift operators to functions on Ω, that is, random variables and sample paths. Throughout this chapter, we assume that random variables and states of stochastic processes take values in a Polish space S with the Borel σ-field $\mathcal{B}(S)$. However, in our applications, it is sufficient to assume that S is a finite dimensional Euclid space, i.e., real valued vector space. As usual, we also assume that a stochastic process is right-continuous with left limits.

Definition 4.3. Let $\{\theta_t\}$ be an operator group on $(\Omega, \mathcal{F})$, and let X be a random variable on this measurable space. Define random variable $X \circ \theta_t$ as

$$X \circ \theta_t(\omega) = X(\theta_t(\omega)), \qquad \omega \in \Omega.$$

With this notation, a stochastic process $\{X(t)\}$ defined on $(\Omega, \mathcal{F})$ is said to be consistent with θ_t if the following condition is satisfied.

$$X(s) \circ \theta_t = X(s+t), \qquad s,t \in \mathbb{R}$$

Similarly, we define the consistency of a discrete time process $\{X_n\}$ with respect to a discrete time shift operator η_n by

□ $$X_m \circ \eta_n = X_{m+n}, \qquad m,n \in \mathbb{Z}.$$

The following definitions of stationary processes are standard.

Definition 4.4. A stochastic process $\{X(t)\}$ is said to be stationary under P if , for each fixed $n \geq 1, t_i \in \mathbb{R}, B_i \in \mathcal{B}(S)$ for $i = 1,2,\ldots,n$,

$$P(X(t_i + u) \in B_i, i = 1,2,\ldots,n)$$

is unchanged for all $u \in \mathbb{R}$. Similarly, the stationarity of a discrete time process $\{X_n\}$ under P_0 is defined, where P_0 is another probability measure on $(\Omega, \mathcal{F})$. □

The next lemma is immediate from the definitions of the shift operators, the consistency and the stationarity.

Lemma 4.1. If P is a θ_t-stationary probability measure and if $\{X(t)\}$ is consistent with $\{\theta_t\}$, then $\{X(t)\}$ is a continuous time stationary process under P. Similarly, if P_0 is η_n-stationary and if $\{X_n\}$ is consistent with $\{\eta_n\}$, then $\{X_n\}$ is a discrete time stationary process under P_0.

It should be noted that we are concerned with different probability measures in Lemma 4.1, but the underlying measurable spaces, i.e., the sample space and the set of all events, are the same. This allows us to directly relate $X(t)$ to X_n through $\omega \in \Omega$.

Example 4.2. How one can create a sample space Ω with operations θ_t and η_n for a queueing model ? Let us consider this problem by a small example. Since an actual system usually starts at some fixed time, we assume that a queueing system starts with no customer at time $c_0 \equiv 0$. Assume that this system is closed with no customer just before time c_1. We represents the evolution of this system on the time interval by a function f from $[c_0, c_1)$ to S, where S is a finite dimensional real vector space. At time c_1, the system restarts and repeats the same trajectory until time $c_2 \equiv 2c_1$. If the system operates in this manner continuously, then we have a trajectory ω_0^+:

$$\omega_0^+(t) = \sum_{n=1}^{\infty} f(t - c_{n-1}) 1(c_{n-1} \le t < c_n), \qquad t \ge 0,$$

where $c_n = nc_1$, and $1(\cdot)$ is the indicator function of the statement "$\cdot$", i.e., it takes 1 (or 0) if the statement is true (or false). We next shift the starting time c_0 to $-kc_1$ for positive integer k, and letting k to infinity, we have the double sided trajectory ω_0:

$$\omega_0(t) = \sum_{n=-\infty}^{+\infty} f(t - c_{n-1}) 1(c_{n-1} \le t < c_n), \qquad t \ge 0.$$

Let $\omega_0^{(u)}(t) = \omega_0(t-u)$ for $u \in [0, c_1)$, and define the sample space Ω as

$$\Omega = \{\omega_0^{(u)}; u \in [0, c_1)\}.$$

Since this sample space is the set of functions on $\mathbb{R}$ and closed under time shift, we have a natural shift operator θ_t. Furthermore, let

$$\eta_n \circ \omega_0^{(u)}(t) = \omega_0(t - c_n), \qquad u \in [0, c_1), t \in \mathbb{R}, n \in \mathbb{Z}.$$

Then, $\{\eta_n\}$ is a discrete time shift operator group. Obviously, this operator group is stationary for any probability measure. Since f is a deterministic function, a probability measure on $(\Omega, \mathcal{F})$ can be determined by that on $[0, c_1) \times \mathcal{B}([0, c_1))$. In particular, if this distribution is uniform on $[0, c_1)$, then P is stationary with respect to the natural shift operator group $\{\theta_t\}$. □

This example is trivial in the sense that all sample paths are generated by a single function $\{\omega_0(t)\}$. Nevertheless, it can be used a prototype of the probability space for shift operators. For example, if we change $c_n - c_{n-1}$ to be *i.i.d.* (that is, independently and identically distributed) random variables and functions on the intervals $[c_{n-1}, c_n)$ to be also *i.i.d.* random functions, then, using the same uniform distribution, we can construct the probability measure which is stationary with respect to the natural shift operator group $\{\theta_t\}$. This construction will be systematically studied in the following two sections.

4.3 Point processes

We introduce a process for randomly chosen discrete time instants on the time axis. This process is called a point process, and will be used to generate a discrete time process, called embedded process, from a continuous time process. Thus, the point process will make a bridge between continuous time and discrete time embedded processes.

Definition 4.5. *N* is called a point process on the line if it satisfies the following two conditions.

(4.3a) N is an integer-valued and locally finite random measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, that is, each $\omega \in \Omega$, $N(\cdot)(\omega)$ is an integer-valued measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $N(B)(\omega) < \infty$ for any bounded $B \in \mathcal{B}(\mathbb{R})$ and $\omega \in \Omega$.

(4.3b) For all $n \geq 1$, $B_i \in \mathcal{B}(\mathbb{R})$ and $n_i \in \mathbb{Z}_+ \equiv \{0,1,\ldots\}$ for $i = 1,2,\ldots,n$,

$$\{N(B_i) = n_i, i = 1,2,\ldots,n\} \in \mathcal{F}.$$

Furthermore, if $N(\{t\}) \leq 1$ for all $t \in \mathbb{R}$, then N is said to be simple.

If we remove the assumption that $N(B)$ is integer-valued, we can similarly define a random measure, but we do not need this generality in this chapter except for Section 4.11.

Similar to the case of a stochastic process, we define the operation of θ_t to point process N as

$$N(B) \circ \theta_t(\omega) = N(B)(\theta_t(\omega)), \qquad t \in \mathbb{R}, \omega \in \Omega, B \in \mathcal{B}(\mathbb{R}).$$

In what follows, we assume

$$N(B) \circ \theta_t = N(t+B), \qquad t \in \mathbb{R}, B \in \mathcal{B}(\mathbb{R}),$$

where $t + B = \{t + u; u \in \mathbb{R}\}$. In this case, N is said to be consistent with θ_t. This is meant that N and θ_t have a common time axis similar to the case of a stochastic process.

The stationarity of point process N is defined similar to that of a stationary process. That is, N is said to be stationary if, for all $n \geq 1$, $k_1,\ldots,k_n \in \mathbb{Z}_+$ and $B_1,\ldots,B_n \in \mathcal{B}(\mathbb{R})$

$$P(N(t+B_1) = k_1, N(t+B_2) = k_2, \ldots, N(t+B_n) = k_n)$$

is unchanged for all $t \in \mathbb{R}$.

The next lemma is a point process version of Lemma 4.1.

Lemma 4.2. If P is θ_t-stationary and if N is consistent with θ_t, then N is stationary under P.

Note that $N((0,t])$ with $t > 0$ can be considered to be a counter for random events that occur in the time interval $(0,t]$. Because of this, a point process is also called a counting process. From this viewpoint, it may be convenient to define the time when the discrete events occur. Let

$$T_n = \begin{cases} \inf\{t > 0; N((0,t]) \geq n\}, & n \geq 1, \\ \sup\{t \leq 0; N((t,0]) \geq 1-n\}, & n \leq 0. \end{cases}$$

This T_n is said to be the n-th counting time of N. Since $T_n = T_{n+1}$ may occur for $n \neq 0$, T_n may not be strictly increasing in n. We thus have

$$\ldots \leq T_0 \leq 0 < T_1 \leq \ldots \tag{4.1}$$

From the definition of T_n, we have

$$N(B) = \sum_{n=-\infty}^{+\infty} 1(T_n \in B), \qquad B \in \mathcal{B}(\mathbb{R}),$$

and the right-hand side of this equation can be written as

$$\int_{-\infty}^{+\infty} 1(u \in B) N(du).$$

Remind that $1(\cdot)$ is the indicator function of the statement "·" (see its definition in Example 4.2).

In the remaining part of this section, we assume that N is a simple point process. In this case, T_n is strictly increasing in n. For convenience, let

$$N(t) = \begin{cases} N((0,t]), & t > 0, \\ -N((t,0]), & t \leq 0. \end{cases}$$

Since N is simple, $N(T_n) = n$ for $n \geq 1$. Note that N is assumed to be consistent with the shift operator θ_t. This yields, for $n \geq 1$ and $s > 0$,

$$\begin{aligned} T_n \circ \theta_s &= \inf\{t > 0; N \circ \theta_s((0,t]) = n\} \\ &= \inf\{t > 0; N((s,s+t]) = n\} \\ &= T_{N(s)+n} - s. \end{aligned} \tag{4.2}$$

For $s \leq 0$ and $n \leq 0$, we can get the same formula (4.2). In particular, letting $s = T_m$ in (4.2), $N(s) = m$ yields

$$T_n \circ \theta_{T_m} = T_{m+n} - T_m.$$

Thus, θ_{T_n} shifts the counting number. From this observation, we define η_n for $n \geq 1$ as

$$\eta_n(\omega) = \theta_{T_n(\omega)}(\omega), \qquad \omega \in \Omega. \tag{4.3}$$

Lemma 4.3. For simple point process N, $\{\eta_n; n \in \mathbb{Z}\}$ is a discrete time shift operator group on $(\Omega, \mathcal{F})$.

Proof. From (4.3), we have

$$\begin{aligned} \eta_n \circ \eta_m &= \theta_{T_n \circ \eta_m} \circ \eta_m \\ &= \theta_{T_{m+n} - T_m} \circ \theta_{T_m} \\ &= \theta_{T_{m+n}} = \eta_{m+n}. \end{aligned}$$

Hence, η_n satisfies (4.2d), which corresponds with (ii) of Definition 4.1. To see condition (4.2c), let $\Phi(\omega) = (T_n(\omega), \omega)$ for $\omega \in \Omega$, which is a function from Ω to $\mathbb{R} \times \Omega$. This function is $\mathcal{F}/(\mathcal{B}(\mathbb{R}) \times \mathcal{F})$-measurable. We next let $\varphi((t, \omega)) = \theta_t(\omega)$, which is a $(\mathcal{B}(\mathbb{R}) \times \mathcal{F})/\mathcal{F}$ measurable function from $\mathbb{R} \times \Omega$ to Ω. Hence, $\eta_n = \varphi \circ \Phi$ is $\mathcal{F}/\mathcal{F}$-measurable, which completes the proof. □

Thus, we get the discrete time shift operator η_n from the continuous time shift operator θ_t. The η_n describes the time shift concerning the point process N. The following observation is intuitively clear, but we give a proof since it is a key of our arguments.

Lemma 4.4. Suppose that stochastic process $\{X(t); t \in \mathbb{R}\}$ and simple point process N are consistent with θ_t. Define discrete time process $\{Y_n; n \in \mathbb{Z}\}$ by $Y_n = X(T_n)$ for the counting times $\{T_n\}$ of N. Then, $\{Y_n\}$ is consistent with η_n.

Proof. From (4.3) and the fact that $X(t)$ is consistent with θ_t, we have

$$\begin{aligned} Y_n \circ \eta_m &= X \circ \eta_m(T_n \circ \eta_m) \\ &= X \circ \theta_{T_m}(T_{m+n} - T_m) \\ &= X(T_{m+n} - T_m + T_m) = Y_{m+n}. \end{aligned}$$

Thus, Y_n is indeed consistent with η_n. □

We next add information to the counting times T_n of N. This information is called mark, and the resulted process is called a marked point process. This process is formally defined in the following way. Let N be a simple point process which is consistent with θ_t, and let $\{T_n\}$ be its counting times. Further, let $\{Y_n\}$ be a discrete time process with state space by $\mathcal{K}$, where $\mathcal{K}$ is assumed to be a Polish space. Then, $\Psi \equiv \{(T_n, Y_n)\}$ is called a marked point process, and Y_n is said to be a mark at the n-th point T_n.

Define a random measure M_Ψ on $(\mathbb{R} \times \mathcal{K}, \mathcal{B}(\mathbb{R}) \times \mathcal{B}(\mathcal{K}))$ as

$$M_\Psi(B, C) = \sum_{n=-\infty}^{+\infty} 1(T_n \in B, Y_n \in C), \qquad B \in \mathcal{B}(\mathbb{R}), C \in \mathcal{B}(\mathcal{K})$$

where $\mathcal{B}(\mathcal{K})$ is the Borel σ-field on $\mathcal{K}$. If we have, for all $t \in \mathbb{R}$,

$$M_\Psi(B,C)\circ\theta_t = M_\Psi(B+t,C), \qquad B\in\mathcal{B}(\mathbb{R}), C\in\mathcal{B}(\mathcal{K}),$$

then Ψ is said to be consistent with θ_t. In particular, if $\{Y_n\}$ is consistent with $\eta_n \equiv \theta_{T_n}$, then M_Ψ is consistent with θ_t. In fact, from (4.2), we have

$$\{T_n\circ\theta_t \in B\} = \{T_{N(t)+n}\in B+t\}.$$

On the other hand, $Y_n = Y_0\circ\eta_n$ yields

$$\begin{aligned} Y_n\circ\theta_t(\omega) &= Y_0(\theta_{T_n\circ\theta_t(\omega)}(\theta_t(\omega))) \\ &= Y_0(\theta_{T_{N(t)(\omega)+n}(\omega)}(\omega)) = Y_{N(t)(\omega)+n}(\omega). \end{aligned}$$

Hence, the claim is proved by

$$M_\Psi(B,C)\circ\theta_t = \sum_{n=-\infty}^{+\infty} 1(T_{N(t)+n}\in B+t, Y_{N(t)+n}\in C) = M_\Psi(B+t,C).$$

We also define the stationarity of Ψ similar to N. Namely, if, for all n and $B_i \in \mathcal{B}(\mathbb{R}), C_i\in\mathcal{B}(\mathcal{K})$ $(i=1,2,\ldots,n)$

$$P(M_\Psi(B_1+t,C_1), M_\Psi(B_2+t,C_2),\ldots,M_\Psi(B_n+t,C_n))$$

is unchanged for all $t\in\mathbb{R}$, then Ψ is said to be stationary under P. Similarly to 4.2, we have the following fact, whose proof is left to the reader.

Lemma 4.5. If P is θ_t-stationary and if marked point process Ψ is consistent with θ_t, then Ψ is stationary under P.

4.4 Palm distribution

One may wonder whether P can be θ_t-stationary and η_n-stationary simultaneously. This may look possible, but it is not true. To see this, we consider time shift operations under P assuming that it is θ_t-stationary.

We first note that the distribution of $\{T_n + t; n\in\mathbb{Z}\}$ is unchanged under P for any $t\in\mathbb{R}$ by the θ_t-stationarity. Hence, shifting the time axis does not change the probability measure. We now shift the time axis subject to the uniform distribution on the unit interval $[0,u]$ independently of everything else for a large but fixed number $u>0$. The choice of this u is not essential in the subsequent arguments, but thinking of the large u may be more appearing. The renumbered $\{T_n\}$ still has the same distribution because P is unchanged. Under such time shifting, T_n is changed to T_1 if the time interval $(T_{n-1},T_n]$ contains the origin. Because the longer time interval has more chance to include the origin, $T_1 - T_0$ would be differently distributed from $T_{n+1}-T_n$ for $n\neq 0$, where the numbers n of T_n are redefined after the time shifting. On the other hand, if P is η_n-stationary, then we have

$$(T_1 - T_0) \circ \eta_n = T_{n+1} - T_n,$$

which implies that $T_1 - T_0$ and $T_{n+1} - T_n$ are identically distributed. Hence, it is impossible that P is θ_t-stationary and η_n-stationary simultaneously.

This observation motivates us to introduce a convenient probability measure for η_n.

Definition 4.6. Suppose that P is θ_t-stationary, point process N is consistent with θ_t and has a finite intensity $\lambda \equiv N((0,1)$. Define nonnegative valued set function P_0 on $\mathcal{F}$ as

$$P_0(A) = \lambda^{-1} E\Big(\int_0^1 1_{\theta_u^{-1}(A)} N(du)\Big), \qquad A \in \mathcal{F}, \tag{4.4}$$

where 1_A is the indicator function of set A, i.e., $1_A(\omega) = 1(\omega \in A)$. Note that $1_{\theta_u^{-1}(A)}(\omega) = 1_A(\theta_u(\omega)) = (1_A \circ \theta_u)(\omega)$. Then, it is easy to see that P_0 is a probability measure on $(\Omega, \mathcal{F})$, which is referred to as a Palm distribution concerning N. Note that N is not necessarily simple in this definition. □

Remark 4.1. (4.4) is equivalent to that, for any function f from Ω to $\mathbb{R}$ which is $\mathcal{F}/\mathcal{B}(\mathbb{R})$-measurable and either bounded or nonnegative, the following equation holds.

$$E_0(f) = \lambda^{-1} E\Big(\int_0^1 f \circ \theta_u N(du)\Big), \tag{4.5}$$

where E_0 represents the expectation concerning P_0. □

Let $A = \{T_0 = 0\}$ in (4.2), then, for any $u \in \mathbb{R}$,

$$\begin{aligned}\theta_u^{-1}(A) &= \{\omega \in \Omega; \theta_u(\omega) \in A\} \\ &= \{T_0 \circ \theta_u = 0\} = \{T_{N(u)} = u\}.\end{aligned}$$

Furthermore, since $N(\{u\}) \geq 1$ implies $T_{N(u)} = u$, we have, from (4.4),

$$P_0(T_0 = 0) = \lambda^{-1} E(N((0,1])) = 1.$$

Hence, N has a mass at the origin under P_0. This means that P_0 is a conditional probability measure given $N(\{0\}) \geq 1$.

The following result is a key to relate P_0 to P when P is θ_t-stationary, where P_0 is the Palm distribution concerning N. The formula (4.6) below is referred to as either Campbell's or Mecke's formula in the literature.

Lemma 4.6. Let $\{X(t)\}$ be a nonnegative valued stochastic process, then we have

$$E\Big(\int_{-\infty}^{+\infty} X(u) \circ \theta_u N(du)\Big) = \lambda E_0\Big(\int_{-\infty}^{+\infty} X(u) du\Big). \tag{4.6}$$

Proof. Define a nonnegative random variable f as

$$f = \int_{-\infty}^{+\infty} X(s)ds = \int_{-\infty}^{+\infty} X(s+u)ds.$$

Substituting this into (4.5), we obtain (4.6) through the following computations.

$$\begin{aligned}
\lambda E_0\Big(\int_{-\infty}^{+\infty} X(s)ds\Big) &= E\Big(\int_0^1 \Big(\int_{-\infty}^{+\infty} X(s+u)\circ\theta_u ds\Big)N(du)\Big)\\
&= \int_{-\infty}^{+\infty} E\Big(\int_{-\infty}^{+\infty} 1(0<u<1)X(s+u)\circ\theta_u N(du)\Big)ds\\
&= \int_{-\infty}^{+\infty} E\Big(\int_{-\infty}^{+\infty} 1(0<u<1)X(s+u)\circ\theta_{u+s} N(du+s)\Big)ds\\
&= \int_{-\infty}^{+\infty} E\Big(\int_{-\infty}^{+\infty} 1(0<u-s<1)X(u)\circ\theta_u N(du)\Big)ds\\
&= E\Big(\int_{-\infty}^{+\infty}\int_{u-1}^{u} dsX(u)\circ\theta_u N(du)\Big)\\
&= E\Big(\int_{-\infty}^{+\infty} X(u)\circ\theta_u N(du)\Big),
\end{aligned}$$

where the third equation is obtained using the fact that P is θ_t-stationary. □

It is notable that $\{X(t)\}$ in 4.6 is not necessarily consistent with θ_t, and therefore it is not necessary stationary under P. The essence of (4.6) lies in the shift invariance of P and Lebesgue measure on $\mathbb{R}$.

Example 4.3 (Little's formula). We derive a famous formula due to Little [20] using 4.6. Consider a service system, where arriving customers get service and leave. Let T_n be the n-th arrival time, where T_n is also defined for $n \le 0$. Let N be a point process generated by these T_n, and let θ_t be a shift operator on Ω. We assume that N is consistent with θ_t. Let U_n be the sojourn time of n-th customer in system. We also assume that $\{U_n; n \in \mathbb{Z}\}$ is consistent with η_n defined by (4.3).

Then, the number of customers $L(t)$ in system at time t is obtained as

$$L(t) = \sum_{n=-\infty}^{+\infty} 1(T_n \le t < T_n + U_n).$$

Assume that $L(t)$ is finite for all $t \in \mathbb{Z}$. Let $N(s) = N((0,s])$, then $T_n\circ\theta_s = T_{N(s)+n} - s$, $U_n = U_0\circ\eta_n$ and $\eta_n\circ\theta_s = \theta_{T_{N(s)+n}}$. Hence,

$$L(t)\circ\theta_s = \sum_{n=-\infty}^{+\infty} 1(T_{N(s)+n} \le s+t < T_{N(s)+n} + U_n) = L(s+t),$$

so $\{L(t)\}$ is consistent with θ_t. Assume that P is θ_t-stationary and $\lambda \equiv E(N((0,1]))$ is finite. Thus, $\{L(t)\}$ is a stationary process under P.

Let $X(u) = 1(T_0 \leq -u < T_0 + U_0)$, then we have

$$\int_{-\infty}^{+\infty} X(u)du = U_0,$$

$$\int_{-\infty}^{+\infty} X(u) \circ \theta_u N(du) = \sum_{n=-\infty}^{+\infty} 1(0 \leq -T_n < U_n) = L(0).$$

Hence, 4.6 yields

$$E(L(0)) = \lambda E_0(U_0). \tag{4.7}$$

This is called Little's formula. □

Let $\Psi = \{(T_n, Y_n)\}$ be a marked point process which is consistent with θ_t, and let N be a point process generated by $\{T_n\}$ with a finite intensity $\lambda \equiv E(N(0,1])$. In 4.6, for each fixed $B \in \mathcal{B}(\mathbb{R}), C \in \mathcal{B}(\mathcal{K})$, let

$$X(u) = 1(u \in B, Y_0 \in C), \qquad t > 0.$$

Since

$$\int_{-\infty}^{+\infty} X(u) \circ \theta_u N(du) = \sum_{n=-\infty}^{+\infty} 1(T_n \in B, Y_n \in C),$$

(4.6) yields

$$E(M_\Psi(B,C)) = \lambda |B| E_0(Y_0 \in C), \qquad t > 0, \tag{4.8}$$

where $|B| = \int_B du$, that is, if B is an interval, then $|B|$ is the length of B. From this, we have known that measure $E(M_\Psi(B,C))$ on $(\mathbb{R} \times \mathcal{K}, \mathcal{B}(\mathbb{R}) \times \mathcal{B}(\mathcal{K}))$ is the product of Lebesgue measure and the distribution of Y_0 under P_0.

Another interesting conclusion of 4.6 is the orderliness of a simple point process.

Corollary 4.1. Assume N is a simple point process that is consistent with θ_t and has a finite intensity λ, then

$$\lim_{t \downarrow 0} \frac{1}{t} E(N(T_1, t]; T_1 < t) = 0, \tag{4.9}$$

and therefore

$$\lim_{t \downarrow 0} \frac{1}{t} P(\theta_{T_1}^{-1}(A), T_1 \leq t) = \lambda P_0(A), \qquad A \in \mathcal{F}. \tag{4.10}$$

Proof. Since $N(T_1, t] = \int_0^t 1(u > T_1) N(du)$ for $t > T_1$ and $T_1 \circ \theta_{-u} = T_{-N(-u,0]+1} + u$ for $u > 0$, 4.6 yields

$$\begin{aligned} E(N(T_1,t];t>T_1) &= E\left(\int_0^t (1(u>T_1)\circ\theta_{-u})\circ\theta_u N(du)\right) \\ &= E\left(\int_0^t 1(T_{-N(-u,0]+1}<0)\circ\theta_u N(du)\right) \\ &= \lambda E_0\left(\int_0^t 1(T_{-N(-u,0]+1}<0)du\right). \end{aligned}$$

Since $P_0(T_0=0)=1$, the indicator function $1(T_{-N(-u,0]+1}<0)$ vanishes as $u\downarrow 0$ (see also (4.1)). Hence, dividing both sides of the above equation by t and letting $t\downarrow 0$, we have (4.9) by the mean value theorem of an elementary calculus. To get (4.10), we apply 4.6 for $X(u)=1(0<u\le t)1_A$ for $t\ge 0$ and $A\in\mathfrak{F}$, then

$$\begin{aligned} \lambda t P_0(A) &= \left(\int_0^t 1_A\circ\theta_u N(du)\right) \\ &= P(\theta_{T_1}^{-1}(A),T_1\le t)+E\left(\int_{T_1+}^t 1_A\circ\theta_u N(du)\right). \end{aligned}$$

Dividing both sides by t and letting $t\downarrow 0$, (4.9) yields (4.10), where the plus sign at T_1 in the integral indicates that the lower point T_1 is not included in the integral region. □

Note that (4.10) gives another way to define the Palm distribution P_0. This may be more intuitive, but the limiting operation may not be convenient in addition to the restriction to a simple point process.

4.5 Inversion formula

We next present basic properties of Palm distribution P_0 and to give a formula to get back P from P_0 directly.

Theorem 4.1. Suppose that N is a simple point process which is consistent with θ_t and has a finite and non-zero intensity $\lambda=E(N((0,1])$. If P is θ_t-stationary, then P_0 is η_n-stationary. Hence, $\{Y_n\}$ of 4.4 is a discrete time stationary process under P_0. Furthermore, P is obtained from P_0 by

$$P(A)=\lambda E_0\left(\int_0^{T_1} 1_{\theta_u^{-1}(A)}du\right), \qquad A\in\mathfrak{F}. \tag{4.11}$$

Proof. The first half is obtained if $P_0(\eta_1^{-1}(A))=P_0(A)$ holds. We prove this using the definition of the Palm distribution (4.4). Because $\eta_1=\theta_{T_1}$ and (4.2) implies

$$\theta_{T_1}\circ\theta_u(\omega)=\theta_{T_{N(u)+1}(\omega)-u}(\theta_u(\omega))=\theta_{T_{N(u)+1}}(\omega),$$

we have

$$\theta_u^{-1}(\eta_1^{-1}(A)) = \{\eta_1 \circ \theta_u \in A\} = \{\theta_{T_{N(u)+1}} \in A\}.$$

Applying this to (4.4), we have

$$\begin{aligned} P_0(\eta_1^{-1}(A)) &= \lambda^{-1} E\Big(\sum_{n=1}^{N(1)} 1(\theta_{T_{n+1}} \in A)\Big) \\ &= \lambda^{-1}\left(E\Big(\sum_{n=1}^{N(1)} 1(\theta_{T_n} \in A)\Big) + P(\theta_{T_{N(1)+1}} \in A) - P(\theta_{T_1} \in A)\right). \end{aligned}$$

Since $\theta_{T_1} \circ \theta_1 = \theta_{T_{N(1)+1}}$ and P is θ_t-stationary, we have

$$P(\theta_{T_{N(1)+1}} \in A) = P(\theta_{T_1} \in A).$$

Thus, we get $P_0(\eta_1^{-1}(A)) = P_0(A)$. We next prove (4.11). For this, let

$$X(u) = 1(N((-u,0)) = 0, u > 0)1_{\theta_u^{-1}(A)},$$

then

$$X(u) \circ \theta_u = 1(N((0,u)) = 0, u > 0)1_A = 1(0 < u \le T_1)1_A.$$

Substituting this in the left-hand side of (4.6), we have

$$E\Big(\int_{-\infty}^{+\infty} X(u) \circ \theta_u N(du)\Big) = E(1_A) = P(A),$$

since $N(du)$ has a unit mass at $u = T_1$. On the other hand, the right-hand side of (4.6) becomes

$$\begin{aligned} E_0\Big(\int_{-\infty}^{+\infty} X(u)du\Big) &= E_0\Big(\int_{T_{-1}}^{0} 1_{\theta_u^{-1}(A)} du\Big) \\ &= E_0\Big(\int_{T_{-1}\circ\eta_1}^{0} 1_{(\theta_u\circ\eta_1)^{-1}(A)} du\Big) \end{aligned}$$

since P_0 is η_n-stationary. Note that $T_{-1} \circ \eta_1 = T_0 - T_1$ and $\theta_u \circ \eta_1 = \theta_{u+T_1}$. Hence, changing the integration variable from u to $u + T_1$ in the last term and using the fact that $P_0(T_0 = 0) = 1$, we have (4.11). □

An excellent feature of the definition (4.4) of Palm distribution P_0 is that it computes the conditional distribution given the event with probability zero, using neither limiting operations nor conditional expectation as a Radon-Nikodym derivative of the measure theory.

From (4.11), P is obtained from P_0. In this sense, it is called an inversion formula. Another interpretation of (4.11) is that it represents the time average of the indicator function of A from $T_0 = 0$ to T_1. Since $\{T_n - T_{n-1}\}$ is stationary under P_0, (4.11) is also called a cycle formula.

The next result shows that the inverse of Theorem 4.1 holds.

Theorem 4.2. Suppose that a simple point process N is consistent with θ_t, a measure P_0 on $(\Omega, \mathcal{F})$ satisfies that $0 < E_0(T_1) < \infty$ for $T_1 \equiv \sup\{u > 0; N(0,u) = 0\}$. Let $\lambda = 1/E_0(T_1)$. If P_0 is η_n-stationary, then P defined by (4.11) is a θ_t-stationary probability measure. Furthermore, $E(N((0,1]) = \lambda$ and (4.4) holds for these P_0 and P.

Proof. It is easy to see that P is a probability measure. Let us show $P(\theta_t^{-1}(A)) = P(A)$ for $A \in \mathcal{F}$ and for all $t \in \mathbb{R}$. From the definition (4.11) of P, we have

$$\begin{aligned} P(\theta_t^{-1}(A)) &= \frac{1}{E_0(T_1)} E_0\Big(\int_0^{T_1} 1_{\theta_{t+u}^{-1}(A)} du\Big) \\ &= \frac{1}{E_0(T_1)} E_0\Big(\int_t^{t+T_1} 1_{\theta_u^{-1}(A)} du\Big) \\ &= \frac{1}{E_0(T_1)} E_0\Big(\int_0^{T_1} 1_{\theta_u^{-1}(A)} du + \int_{T_1}^{t+T_1} 1_{\theta_u^{-1}(A)} du - \int_0^{t} 1_{\theta_u^{-1}(A)} du\Big). \end{aligned}$$

Since P_0 is η_n-stationary, we have $E_0(X) = E_0(X \circ \eta_1)$ for a nonnegative random variable X. Hence, we have

$$\begin{aligned} E_0\Big(\int_0^{t} 1_{\theta_u^{-1}(A)} du\Big) &= E_0\Big(\int_0^{t} 1_{(\theta_u \circ \eta_1)^{-1}(A)} du\Big) \\ &= E_0\Big(\int_0^{t} 1_{\theta_{u+T_1}^{-1}(A)} du\Big) \\ &= E_0\Big(\int_{T_1}^{t+T_1} 1_{\theta_u^{-1}(A)} du\Big). \end{aligned}$$

Thus, we get $P(\theta_t^{-1}(A)) = P(A)$, so P is θ_t-stationary. It remains to prove (4.4), where P is defined by (4.11). Using this P, define $P_0^\dagger$ as

$$P_0^\dagger(A) = \frac{1}{E(N((0,1]))} E\Big(\int_0^1 1_{\theta_u^{-1}(A)} N(du)\Big), \qquad A \in \mathcal{F}.$$

Thus, the proof is completed if we show $P_0 = P_0^\dagger$. This equality is equivalent to that, for nonnegative and bounded random variable f,

$$E_0(f) = \frac{1}{E(N((0,1]))} E\Big(\int_0^1 f \circ \theta_u N(du)\Big).$$

From (4.11),

$$E\Big(\int_0^1 f \circ \theta_u N(du)\Big) = \lambda E_0\Big(\int_0^{T_1} \Big(\int_0^1 f \circ \theta_u N(du)\Big) \circ \theta_s ds\Big). \tag{4.12}$$

Hence, we need to verify that

$$E_0(f) = \frac{\lambda}{E(N((0,1]))} E_0\Big(\int_0^{T_1} \Big(\int_0^1 f \circ \theta_u N(du) \Big) \circ \theta_s ds \Big). \tag{4.13}$$

Let us prove (4.13). We first compute the inside of the expectation in the right-hand side of (4.13). Since Lebesgue integration is unchanged by shifting the integration variable, we have

$$\begin{aligned} \int_0^{T_1} \Big(\int_0^1 f \circ \theta_u N(du) \Big) \circ \theta_s ds &= \int_0^{T_1} \Big(\int_0^1 f \circ \theta_{u+s} N(du+s) \Big) ds \\ &= \int_0^{T_1} \Big(\int_s^{s+1} f \circ \theta_u N(du) \Big) ds \\ &= \int_{-\infty}^{+\infty} \int_0^{T_1} 1(s < u < s+1) ds (f \circ \theta_u) N(du) \\ &= \sum_{n=-\infty}^{+\infty} \int_0^{T_1} 1(s < T_n < s+1) ds (f \circ \theta_{T_n}). \end{aligned}$$

Since P_0 is stationary with respect to $\eta_n = \theta_{T_n}$, the expectation concerning P_0 in the left-hand side of (4.13) becomes

$$\begin{aligned} &\sum_{n=-\infty}^{+\infty} E_0\Big(\int_0^{T_1} 1(s < T_n < s+1) ds (f \circ \theta_{T_n}) \Big) \\ &\quad = \sum_{n=-\infty}^{+\infty} E_0\Big(\Big(\int_0^{T_1} 1(s < T_n < s+1) ds f \circ \eta_n \Big) \circ \eta_{-n} \Big) \\ &\quad = \sum_{n=-\infty}^{+\infty} E_0\Big(\int_0^{T_{1-n} - T_{-n}} 1(s < T_0 - T_{-n} < s+1) ds f \Big) \\ &\quad = E_0\Big(\sum_{n=-\infty}^{+\infty} \int_{T_{-n}}^{T_{1-n}} 1(s < T_0 < s+1) ds f \Big) \\ &\quad = E_0\Big(\int_{-\infty}^{+\infty} 1(T_0 - 1 < s < T_0) ds f \Big) = E_0(f). \end{aligned}$$

Thus, (4.13) is obtained if $\lambda = E(N((0,1]))$. The latter is obtained from (4.12) with $f \equiv 1$ and the above computations. This completes the proof. □

In applications, particularly in queueing networks, we frequently meet the situation that point process N is the superposition of m point processes $N_1, \ldots, N_m$ all of which are consistent with θ_t for some $m \geq 2$. Namely,

$$N(B) = N_1(B) + \ldots + N_m(B), \qquad B \in \mathcal{B}(\mathbb{R}).$$

Assume that P is θ_t-stationary, and $\lambda \equiv E(N((0,1]) < \infty$ with $\lambda \neq 0$. For $i = 1, 2, \ldots, m$, let $\lambda_i = E(N_i((0,1]))$, and denote Palm distribution concerning N_i by

P_i. Then, from the definition of Palm distribution, it is easy to see that

$$\lambda P_0(A) = \sum_{i=1}^{m} \lambda_i P_i(A), \qquad A \in \mathcal{F}. \tag{4.14}$$

This decomposition of the Palm measure is shown to be useful in applications (see Section 4.10).

4.6 Detailed Palm distribution

We have mainly considered Palm distribution when point process N is simple. The definition of Palm distribution itself does not need for N to be simple. However, if N is not simple, η_n defined by $\eta_n = \theta_{T_n}$ can not properly handle events that simultaneously occur in time. We need to differently define Palm distribution for this case. To this end, we consider a pair of T_n and η_n, where η_n is a discrete time operator group. In what follows, point process N is assumed to be generated by $\{T_n\}$, and N is not necessarily simple.

Definition 4.7. Let $\{T_n; n \in \mathbb{Z}\}$ be a nondecreasing sequence of random variables, which generate point process N, and let $\{\eta_n\}$ be the discrete time shift operator group. If

$$\{(T_n, \eta_n)\} \circ \theta_t = \{(T_n - t, \eta_n)\}, \qquad n \in \mathbb{Z}, t \in \mathbb{R}$$

holds, then $\{(T_n, \eta_n)\}$ is said to be a θ_t-consistent marked point process with shift operator η_n.

Definition 4.8. Suppose that P be θ_t-stationary and $\{(T_n, \eta_n)\}$ is a θ_t-consistent marked point process with η_n, where $\ldots \le T_{-1} \le T_0 \le 0 < T_1 \le T_2 \le \ldots$, and $\lambda \equiv E(N((0,1])) < \infty$. Then, we define $\overline{P}_0$ as

$$\overline{P}_0(A) = \lambda^{-1} E\Big(\sum_{n=1}^{N((0,1])} 1_{\eta_n^{-1}(A)} \Big), \qquad A \in \mathcal{F}. \tag{4.15}$$

This $\overline{P}_0$ is a probability measure on $(\Omega, \mathcal{F})$, and called a detailed Palm distribution concerning $\{T_n\}$.

Detailed Palm distribution $\overline{P}_0$ is different from Palm distribution P_0 concerning N if N is not simple. Nevertheless, we can extend the results in the previous two sections to detailed Palm distribution. Since their proofs are similar to the previous ones, we present their versions for Theorems 4.1 and 4.2 without proof.

Theorem 4.3. Suppose that P is θ_t-stationary, $\{(T_n, \eta_n)\}$ is a θ_t consistent marked point process, and $\lambda \equiv E(N((0,1])) < \infty$. Then, detailed Palm distribution $\overline{P}_0$ is η_n-stationary. Furthermore, P is recovered from $\overline{P}_0$ by

$$P(A) = \lambda \overline{E}_0\Big(\int_0^{T_1} 1_{\theta_u^{-1}(A)} du\Big), \qquad A \in \mathcal{F}. \tag{4.16}$$

where $\overline{E}_0$ represents the expectation concerning $\overline{P}_0$. Conversely, suppose that probability measure $\overline{P}_0$ on $(\Omega, \mathcal{F})$ satisfies $0 < E_0(T_1) < \infty$ and $\overline{P}_0$ is η_n-stationary for a given discrete time shift operator group $\{\eta_n\}$. Let $\lambda = 1/\overline{E}_0(T_1)$ and define P by (4.16), Then, P is a probability measure on $(\Omega, \mathcal{F})$ which is θ_t-stationary, and we have $E(N((0,1]) = 1/E_0(T_1) = \lambda$. For these $\overline{P}_0$ and P, we have (4.15).

We next consider another way to define the detailed Palm distribution. For this, we use a simple point process which have masses at the same time instant as N. Denote this point process by N^*. Namely, N^* is defined as

$$N^*(B) = \int_B \frac{1}{N(\{u\})} N(du), \qquad B \in \mathcal{B}(\mathbb{R}). \tag{4.17}$$

This point process is said to be a simple version of N. Let T_n^* be the n-th counting point of N^*.

Lemma 4.7. Under the same assumptions of Definition 4.8, let $\lambda^* = E(N^*((0,1]))$ and denote the Palm distribution concerning N^* by P_0^*, then we have

$$\overline{P}_0(A) = \frac{\lambda^*}{\lambda} E_0^* \Big(\sum_{n=1}^{N((0,T_1^*])} 1_{\eta_n^{-1}(A)} \Big), \qquad A \in \mathcal{F}, \tag{4.18}$$

where E_0^* represents the expectation concerning P_0^*.

Proof. Let $\eta_n^* = \theta_{T_n^*}$. Then, from the definition of Palm distribution P_0^*, we have, for random variable f,

$$E_0^*(f) = (\lambda^*)^{-1} E \left(\sum_{n=1}^{N^*((0,1])} f \circ \eta_n^* \right).$$

In this equation, letting $f = \sum_{\ell=1}^{N((0,T_1^*])} 1_{\eta_\ell^{-1}(A)}$, the expectation in the right-hand side becomes

$$\begin{aligned}
E\left(\sum_{n=1}^{N^*((0,1])}\left(\sum_{\ell=1}^{N((0,T_1^*])} 1_{\eta_\ell^{-1}(A)}\right)\circ\eta_n^*\right)
&= E\left(\sum_{n=1}^{N^*((0,1])}\left(\sum_{\ell=1}^{N((T_n^*,T_{n+1}^*])} 1_{(\eta_{N((0,T_n^*])+\ell})^{-1}(A)}\right)\right)\\
&= E\left(\sum_{n=1}^{N^*((0,1])}\left(\sum_{\ell=N((0,T_n^*])+1}^{N((0,T_{n+1}^*])} 1_{(\eta_\ell)^{-1}(A)}\right)\right)\\
&= E\left(\sum_{n=1}^{N((0,1])} 1_{\eta_\ell^{-1}(A)}\right).
\end{aligned}$$

Since the last term equals $\lambda\overline{P}_0(A)$ by (4.15), we have (4.18). □

Example 4.4. Let us consider batch arrival queueing system. Let T_n^* be the n-th batch arrival time. We then number all customers sequentially including those who are in the same batch. Let T_n be the n-th arrival time of a customer in this sense. Let B_n the size of the batch arriving at time T_n^*, and let J_n be the number of the n arriving customer counted in his batch. That is, $J_n = \max\{\ell \geq 1; T_n = T_{n-\ell+1}\}$. In particular, $J_0 = B_0$. Let $\eta_n = \theta_{T_n}$, then

$$J_0 \circ \eta_n = \max\{\ell \geq 1; 0 = T_{n-\ell+1} - T_n\} = J_n.$$

Hence, if $\{T_n\}$ is η_n-stationary under $\overline{P}_0$, then we have, for any $n \in \mathbb{Z}$,

$$\begin{aligned}
\overline{P}_0(J_n = k) &= \overline{P}_0(J_0 = k)\\
&= \frac{\lambda^*}{\lambda}E_0^*\Big(\sum_{n=1}^{B_1} 1(J_0 \circ \eta_n = k)\Big) = \frac{1}{E_0^*(B_1)}P_0^*(B_1 \geq k),
\end{aligned}$$

because $J_n = k$ for some positive $n \leq B_1$ if and only if $B_1 \geq k$. This means that a randomly chosen customer is counted in its batch subject to the so called stationary excess distribution of B_1 under P_0^*. □

4.7 Time and event averages

In this section, we give interpretations of stationary P and P_0 through sample averages. It will be shown that these sample averages are unchanged under both of them. This means that both probability measures can be used for computing stationary characteristics when either one of them is taken for a probability model. Furthermore, sample averages may be only a way to identify system parameters. Thus, the unchanged sample averages are particularly important in applications of Palm calculus. This is something like to use two machines for production which is originally designed for one machine. Throughout this section, we assume

(4.7a) Measurable space $(\Omega, \mathcal{F})$ is equipped with a shift operator group $\{\theta_t; t \in \mathbb{R}\}$.

(4.7b) There exists a simple point process N which is consistent with θ_t, and the discrete time shift operator group $\{\eta_n; n \in \mathbb{R}\}$ is defined by (4.3).

(4.7c) There exists a probability measure P on $(\Omega, \mathcal{F})$ which is θ_t-stationary and satisfies $\lambda \equiv E(N(0,1]) < \infty$.

By these assumptions, Palm distribution P_0 is well defined for N. Let

$$\mathcal{I} = \{A \in \mathcal{F}; \theta_t^{-1}(A) = A \text{ holds for all } t \in \mathbb{R}\},$$

then $\mathcal{I}$ is σ-field on Ω. Since $\theta_t^{-1}(\mathcal{I}) = \mathcal{I}$, this $\mathcal{I}$ is called an invariant σ-field concerning θ_t. Similarly, an invariant σ-field concerning η_n is defined.

Lemma 4.8. For the shift operator group $\{\eta_n; n \in \mathbb{Z}\}$, define $\mathcal{I}_0$ as

$$\mathcal{I}_0 = \{A \in \mathcal{F}; \eta_1^{-1}(A) = A\}.$$

Then, $\mathcal{I}_0 = \mathcal{I}$, and $\mathcal{I}_0$ is the invariant σ-field concerning η_n.

Proof. From the definition, $\mathcal{I}_0$ is clearly η_n-invariant, i.e., $\eta_n^{-1}(\mathcal{I}_0) = \mathcal{I}_0$, Hence, we only need to prove $\mathcal{I}_0 = \mathcal{I}$. Choose $A \in \mathcal{I}$. Since $\theta_t^{-1}(A) = A$, we have

$$\begin{aligned}\eta_1^{-1}(A) &= \{\omega \in \Omega; \theta_{T_1(\omega)}(\omega) \in A\} \\ &= \cup_{t \in \mathbb{R}} \{T_1 = t\} \cap \theta_t^{-1}(A) \\ &= \cup_{t \in \mathbb{R}} \{T_1 = t\} \cap A = A.\end{aligned}$$

Thus, we have $A \in \mathcal{I}_0$. Conversely, let $A \in \mathcal{I}_0$. Since $\eta_n \circ \eta_1 = \eta_{n+1}$, we have $\eta_n^{-1}(A) = A$ for any $n \in \mathbb{Z}$. If $T_{n-1} \le t < T_n$, then

$$\eta_1 \circ \theta_t(\omega) = \theta_{T_1(\theta_t)}(\theta_t(\omega)) = \theta_{T_n(\omega)-t}(\theta_t(\omega)) = \theta_{T_n}(\omega) = \eta_n(\omega).$$

Hence, for any $t \in \mathbb{R}$,

$$\begin{aligned}\theta_t^{-1}(A) &= \cup_{n=-\infty}^{+\infty} \{T_{n-1} \le t < T_n\} \cap \theta_t^{-1}(A) \\ &= \cup_{n=-\infty}^{+\infty} \{T_{n-1} \le t < T_n\} \cap \theta_t^{-1}(\eta_1^{-1}(A)) \\ &= \cup_{n=-\infty}^{+\infty} \{T_{n-1} \le t < T_n\} \cap (\eta_1 \circ \theta_t)^{-1}(A) \\ &= \cup_{n=-\infty}^{+\infty} \{T_{n-1} \le t < T_n\} \cap \eta_n^{-1}(A) \\ &= \cup_{n=-\infty}^{+\infty} \{T_{n-1} \le t < T_n\} \cap A = A.\end{aligned}$$

Thus, we have $A \in \mathcal{I}$, which completes the proof. □

For $A \in \mathcal{I}$, $P(A) \neq P_0(A)$ in general, but we have the following result.

Lemma 4.9. For $A \in \mathcal{I}$, $P_0(A) = 1$ if and only if $P(A) = 1$.

Proof. Since $\theta_u^{-1}(A) = A$ for $A \in \mathcal{I}$, from (4.4) and (4.11), it follows that

$$P_0(A) = \lambda^{-1} E(1_A N((0,1])), \qquad P(A) = \frac{1}{E_0(T_1)} E_0(1_A T_1).$$

Hence, if $P_0(A) = 1$, then $E(1_A N((0,1])) = E(N((0,1])$, which implies $P_0(A) = 1$. Conversely, if $P(A) = 1$, then $E_0(1_A T_1) = E_0(T_1)$, which implies $P(A) = 1$. □

Clearly, the equivalence in this lemma is not true for $A = \{T_0 = 0\}$. Thus, it may not be true for $A \notin \mathfrak{I}$.

Definition 4.9. Suppose that probability measure P on $(\Omega, \mathcal{F})$ is θ_t-stationary, and let $\mathfrak{I}$ be the invariant σ-field concerning θ_t. If either $P(A) = 0$ or $P(A) = 1$ for each $A \in \mathfrak{I}$, then P is said to be ergodic concerning θ_t. As for P_0 and $\{\eta_n; n \in \mathbb{Z}\}$, we similarly define P_0 to be ergodic concerning η_n.

From 4.9, the following result is immediate.

Lemma 4.10. Assume that probability measure P on $(\Omega, \mathcal{F})$ is θ_t-stationary. Then, P_0 is ergodic concerning η_n if and only if P is ergodic concerning θ_t.

The next result is a version of law of large numbers, and called ergodic theorem. We omit its proof, which can be found in text books on probability theory (see, e.g., [5]).

Theorem 4.4. Let $\{\eta_n; n \in \mathbb{R}\}$ be the shift operator group on $(\Omega, \mathcal{F})$, and let $\{Y_n\}$ be a discrete time stochastic process which is consistent with η_n. Let P_0 be a η_n-stationary probability measure on $(\Omega, \mathcal{F})$, and denote the expectation concerning P_0 by E_0. If $E_0(|Y_0|) < \infty$, then we have, under P_0,

$$\lim_{n \to \infty} \frac{1}{n} \sum_{\ell=1}^{n} Y_\ell = E_0(Y_0 | \mathfrak{I}_0) \tag{4.19}$$

with probability one, where $\mathfrak{I}_0$ is the invariant σ-field concerning η_n, and $E_0(Y_0|\mathfrak{I}_0)$ is the conditional expectation of Y_0 given $\mathfrak{I}_0$.

This theorem leads to the following results.

Corollary 4.2. Suppose (4.7a), (4.7b), (4.7c). Then, $\{Y_n\}$ is consistent with η_n. If $E_0(|Y_0|) < \infty$, then we have, under both of P_0 and P,

$$\lim_{n \to \infty} \frac{1}{n} \sum_{\ell=1}^{n} Y_\ell = E_0(Y_0 | \mathfrak{I}) \tag{4.20}$$

with probability one. Furthermore, $\{X(t)\}$ is consistent with θ_t, and if $E(|X_0|) < \infty$, then we have, under both of P_0 and P,

$$\lim_{t \to \infty} \frac{1}{t} \int_0^t X(u) du = E(X(0) | \mathfrak{I}) \tag{4.21}$$

with probability one.

Remark 4.2. Sample averages in (4.20) and (4.21) are referred to as event and time averages, respectively. If P or P_0 is ergodic, then $\mathfrak{I}$ consists of events which have either probability zero or probability one. Hence, the conditional expectations in (4.20) and (4.21) reduce to the unconditional ones. If Y_n (or $X(t)$) is nonnegative, we do not need the condition that $E_0(Y_0) < \infty$ (or $E(X(0)) < \infty$). To see this, we first apply $\min(a,X(t))$ to (4.20) for fixed constant $a > 0$, then let $a \uparrow \infty$. □

Proof. Since P_0 is η_n-stationary, (4.20) holds under P_0 with probability one by Theorem 4.4. Let A be the set of all $\omega \in \Omega$ such that (4.20) holds. Since $P_0(A) = 1$, 4.9 yields $P(A) = 1$. Hence, (4.20) holds under P with probability one. As for (4.21), if it holds under P with probability one, we similarly get it under P_0. To get (4.21) under P, we let

$$\eta_n = \theta_n, \qquad Y_n = \int_{n-1}^{n} X(u)du.$$

Then, it can be shown that Theorem 4.4 yields (4.21) under P since P is also stationary concerning this η_n. □

Similarly to this corollary, we can prove the next result.

Corollary 4.3. Under the same assumptions of 4.2, we have, under both of P_0 and P,

$$\lim_{t\to\infty} \frac{N((0,t])}{t} = \lim_{t\to\infty} \frac{N((-t,0])}{t} = E(N((0,1])|\mathfrak{I}) \tag{4.22}$$

holds with probability one.

4.2 and 4.3 are convenient to compute sample averages since we can choose either P or P_0 to verify them for both of P and P_0.

Example 4.5 (Little's formula in sample averages). We consider the same model discussed in Example 4.3. Consider a service system, where arriving customers get service and leave. For simplicity, we here assume that all T_n are distinct, i.e., not more than one customers arrive at once. If either P or P_0 is ergodic, then, by 4.2, we can rewrite Little's formula (4.7) in terms of time and event averages as

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t L(u)du = \lambda \lim_{n\to\infty} \frac{1}{n} \sum_{\ell=1}^{n} U_n,$$

which holds with probability one under both of P and P_0.

The simplicity condition on N is not essential in the above arguments. We only need to replace E_0 by the expectation $\overline{E}_0$ of the detailed Palm distribution. □

4.8 Rate conservation law

In the previous sections, we have considered two kinds of expectations by a stationary probability measure and its Palm distributions. Computations using those distributions is called Palm calculus. This calculus gives relationship among characteristics observed at arbitrary points in time and those in embedded epochs. Typical formulas are (4.4), (4.6) and (4.11). They can be applied to a stochastic process. However, they are generally not so convenient for studying a complex systems such as queueing networks. Dynamics of those systems is typically driven by differential operators such as generators and transition rate matrices of Markov processes or chains while formulas in Palm calculus concern integrations over time in general.

In this section, we consider a convenient form of Palm calculus for stochastic processes. As we shall see, this form can be used to characterize the stationary distribution when they are Markov processes. However, in this section, we do not assume any Markovian assumption, but use the same framework as Palm calculus. So, our assumptions is basically of the stationary of processes. Extra assumptions that we need is the smoothness of a sample path except jump instants, which is not so restrictive in queueing applications.

Throughout this section, we assume (4.7a), (4.7b), (4.7c) of Section 4.7. Since these assumptions are important in our arguments, we restate it as follows.

(4.8a) There is a probability space $(\Omega, \mathcal{F}, P)$ such that shift operator group $\{\theta_t; t \in \mathbb{R}\}$ is defined on Ω and P is θ_t-stationary. There is a simple point process N which is consistent with θ_t and satisfies $\lambda \equiv E(N(0,1]) < \infty$.

We further assume the following three conditions on a stochastic process of interest.

(4.8b) $\{X(t)\}$ is a real valued continuous time stochastic process such that it is consistent with θ_t and right-continuous with left-limits for each $t \in \mathbb{R}$, that is, $\lim_{\varepsilon \downarrow 0} X(t+\varepsilon)(\omega) = X(t)(\omega)$, and $X(t-)(\omega) \equiv \lim_{\varepsilon \downarrow 0} X(t-\varepsilon)(\omega)$ exists for each $t \in \mathbb{R}$ and each $\omega \in \Omega$.

(4.8c) At all $t, X(t)$ has the right-hand derivative $X'(t)$. That is,

$$X'(t) \equiv \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon}(X(t+\varepsilon) - X(t))$$

exists and finite.

(4.8d) N includes all the times when $X(t)$ is discontinuous in t. That is, for each $B \in \mathcal{B}(\mathbb{R}), N(B) = 0$ implies $\sum_{t \in B} 1(X(t) \neq X(t-)) = 0$.

All these conditions are sufficient to hold with probability one for our arguments. However, we rather prefer that they hold for all $\omega \in \Omega$ for simplicity.

Lemma 4.11. Under assumptions (4.8a), (4.8b), (4.8c) and (4.8d), $\{X(t)\}$ and $\{X'(t)\}$ are stationary processes, and N is a stationary simple point process. If $E(X'(0))$ and $E_0(X(0-)) - X(0))$ are finite, then

$$E(X'(0)) = \lambda E_0(X(0-) - X(0)). \tag{4.23}$$

Proof. From (4.8a) and (4.8b), $X(t)$ and N are clearly stationary. From the consistency on $X(t)$ and the differentiability (4.8c), it follows that

$$\begin{aligned} X'(t)\circ\theta_u &= \lim_{\varepsilon\downarrow 0}\frac{1}{\varepsilon}(X(t+\varepsilon)-X(t))\circ\theta_u \\ &= \lim_{\varepsilon\downarrow 0}\frac{1}{\varepsilon}(X(t+u+\varepsilon)-X(t+u)) = X'(t+u). \end{aligned}$$

Hence, $X'(t)$ is also consistent with θ_t, so $\{X'(t)\}$ is a stationary process. From (4.8a), $\lambda<\infty$, so $N(0,1]$ is finite with probability one. Hence, from (4.8d), we have

$$X(t) = X(0) + \int_0^t X'(u)du + \int_0^t (X(0)-X(0-))\circ\theta_u N(du), \qquad t>0. \quad (4.24)$$

We tentatively suppose that $E(X(t))$ is finite, which implies that $E(X(t)) = E(X(0))$ due to the stationarity of $X(t)$. Since $E(X'(0))$ is finite and $X'(t)$ is stationary, we have

$$E\Big(\int_0^1 X'(u)du\Big) = \int_0^1 E(X'(u))du = E(X'(0)).$$

Hence, taking the expectations of (4.24) for $t=1$, we have

$$E(X'(0)) + E\Big(\int_0^1 (X(0)-X(0-))\circ\theta_u N(du)\Big) = 0.$$

This yields (4.23) by applying the Palm calculus in Definition 4.6.

We next remove the assumption that $E(X(t))$ is finite. To this end, for each integer $n\geq 1$, define function f_n as

$$f_n(x) = \begin{cases} x, & |x|\leq n, \\ -\frac{1}{2}(\max(0,n+1-x))^2 + \frac{1}{2} + n, & x>n, \\ \frac{1}{2}(\max(0,n+1+x))^2 - \frac{1}{2} - n, & x<-n. \end{cases}$$

This function is bounded since $|f_n(x)|\leq \frac{1}{2}+n$ for all $x\in\mathbb{R}$. Furthermore, it has the right-hand derivative:

$$f_n'(x) = \begin{cases} 1, & -n\leq x<n, \\ n+1-x, & n\leq x<n+1, \\ n+1+x, & -n-1\leq x<-n, \\ 0, & \text{otherwise}. \end{cases}$$

From this, it is easy to see that $f'(x)$ is continuous in x, and $|f_n'(x)|\leq 1$. Furthermore,

$$|f_n(x)-f_n(y)| \leq \int_y^x |f_n'(z)|dz \leq |x-y|.$$

Let $Y_n(t) = f_n(Y(t))$, then $Y_n(t)$ is bounded and

$$|Y_n'(t)| = |Y'(t)f_n'(Y(t))| \leq |Y'(t)|,$$
$$|Y_n(t-) - Y_n(t)| \leq |Y(t-) - Y(t)|,$$

so $E(Y_n'(0))$ and $E_0(Y_n(0-) - Y_n(0))$ are finite by the assumptions. Hence, from the first part of this proof, (4.23) is obtained for $Y_n(t)$. Namely, we have

$$E(Y_n'(0)) = \lambda E_0(Y_n(0-) - Y_n(0)).$$

Let $n \to \infty$ in this equation noting that $f_n(x) \to x$ and $f_n'(x) \uparrow 1$ as $n \to \infty$. Then, the bounded convergence theorem yields (4.23) since $|Y_n'(t)|$ and $|Y_n(t-) - Y_n(t)|$ are uniformly bounded in n. □

Remark 4.3. From the proof of 4.11, we can see that, if $E(X(0))$ is finite, then the finiteness of either $E(X'(0))$ or $E_0(X(0-) - X(0))$ is sufficient to get (4.23). Here, we note that the finiteness of $E(X)$ for a random variable X is equivalent to the finiteness of $E(|X|)$ due to the definition of the expectation.

Formula (4.23) is referred to as a rate conservation law, RCL for short. In fact, it can be interpreted that the total rate due to continuous and discontinuous changes of $X(t)$ are kept zero. In application of the RCL, $X(t)$ is a real or complex valued function of a multidimensional process. Let $\mathbf{X}(t) = (X_1(t),\ldots,X_d(t))$ be such a process for a positive integer d, and let f be a partially differentiable function from $\mathbb{R}^d$ to $\mathbb{R}$. In this case, we put $X(t) = f(\mathbf{X}(t))$. Then, 4.11 yields

Corollary 4.4. Let d be a positive integer, and let f be a continuously partially differentiable function from $\mathbb{R}^d$ to $\mathbb{R}$. If each $X_\ell(t)$ instead of $X(t)$ satisfies conditions (4.8a), (4.8b), (4.8c) and (4.8d) for all $\ell = 1,2,\ldots,d$ and if $E(f(\mathbf{X}(0)))$ and $E(f(\mathbf{X}(0)) - f(\mathbf{X}(0-)))$ are finite, then we have

$$E\left(\mathbf{X}'(0)\nabla f(\mathbf{X}(0))\right) = \lambda E_0(f(\mathbf{X}(0-)) - f(\mathbf{X}(0))). \tag{4.25}$$

where $\mathbf{X}'(0) = (X_1'(0),\ldots,X_d'(0))$ and $\nabla f(\mathbf{x}) = (\frac{\partial}{\partial x_1}f(\mathbf{x}),\ldots\frac{\partial}{\partial x_d}f(\mathbf{x}))^{\mathrm{T}}$ for $\mathbf{x} = (x_1,\ldots,x_d)$.

This is the most convenient form for queueing applications even for $d = 1$ because we can choose any f as far as it is differentiable and satisfies the finiteness conditions on the expectations. We refer this type of f as a test function.

Example 4.6 (Workload process). We consider the workload process with a state dependent processing rate r and an input generated by a point process N and a sequence of input works $\{S_n\}$. The workload $V(t)$ at time $t \geq 0$ is defined as

$$V(t) = V(0) + \sum_{n=1}^{N(t)} S_n - \int_0^t r(V(u))1(V(u) > 0)du,$$

where $r(x)$ is a nonnegative valued right-continuous function on $[0,\infty)$. If $r(x) \equiv 1$, then $V(t)$ is the workload process of a single server queue.

Let T_n be the n-th point of N. We assume that N has a finite intensity λ, $\{(T_n,S_n)\}$ is consistent with θ_t (see Definition 4.7), and $\{V(t)\}$ is a stationary process under P. Let f be a bounded and continuously differentiable function on $\mathbb{R}$. Since $X(t) \equiv V(t)$ satisfies all the conditions of 4.4 and $X'(t) = r(t)$, we have

$$E(r(V(0))f'(V(0))1(V(0) > 0)) = \lambda E_0(f(V(0-)) - f(V(0-)+S_0)). \tag{4.26}$$

Using 4.4, we can generalize this formula for a multidimensional workload process with a multidimensional input. The virtual waiting time vector of a many server queue is such an example. □

Another useful form is obtained from decomposing the point process N.

Corollary 4.5. Under the assumptions of 4.11, suppose that the point process N is decomposed into m point processes $N_1, N_2, \ldots, N_m$ all of which are consistent with θ_t for some $m \geq 2$. Namely,

$$N(B) = N_1(B) + N_2(B) + \ldots + N_m(B), \qquad B \in \mathcal{B}(\mathbb{R}).$$

Further suppose that $\lambda_i \equiv E(N_i((0,1]))$ is finite for all $i = 1,2,\ldots,m$, and denote the expectation concerning Palm distribution with respect to N_i by E_i. If $E_i(X(0)), E_i(X(0))$ are finite for $i = 1,2,\ldots,m$ and if $E(X'(0))$ is finite, then we have

$$E(X'(0)) = \sum_{i=1}^{m} \lambda_i E_i(X(0-) - X(0)). \tag{4.27}$$

Proof. From the assumption on the finiteness of λ_i, N has the finite intensity $\lambda \equiv \sum_{i=1}^{m} \lambda_i$. Denote the expectation concerning Palm distribution with resect to N by E_0, then

$$\begin{aligned}\lambda E_0(X(0-) - X(0)) &= E\left(\int_0^1 (X(u-) - X(u))N(du)\right)\\ &= \sum_{i=1}^{m} E\left(\int_0^1 (X(u-) - X(u))N_i(du)\right)\\ &= \sum_{i=1}^{m} \lambda_i E_i(X(0-) - X(0)).\end{aligned}$$

Hence, 4.11 concludes (4.27). □

We have been only concerned with the simple point process N for the rate conservation law (4.23). If N is not simple, we can use its simple version N^* defined by (4.17). However, we must be careful about the changes of $X(t)$ at atoms of N^*. That is, we need to use the detailed Palm distribution $\overline{P}_0$ instead of P_0 in (4.23) when we consider embedded events at T_n.

4.9 PASTA: a proof by the rate conservation law

In queueing problems, we frequently require to compute system characteristics observed at different points in time. In this section, we demonstrate how we can use the rate conservation law to the observation of a customer arriving subject to a Poisson process, where point process N is called a Poisson process if $\{t_{n+1} - t_n; n \in \mathbb{Z}\}$ is the sequence of independently, identically and exponential random variables for the n-th increasing instants instant t_n of N. The following result is called PASTA, which is the abbreviation of "Poisson Arrivals See Time Averages", which is coined by Wolff [35].

Theorem 4.5 (PASTA). Under the assumptions (4.8a), (4.8b), (4.8c) and (4.8d), let N_0 be the Poisson process which is consistent with θ_t and finite intensity λ_0, and denote the expectation of Palm distribution with respect to N_0 by E_0. If $\{X(u); u < t\}$ is independent of $\{N_0([t, t+s]); s \geq 0\}$ for all t, then we have, for all measurable function f such that $E(f(X(0)))$ and $E_0(f(X(0-)))$ are finite,

$$E(f(X(0))) = E_0(f(X(0-))). \tag{4.28}$$

Proof. Similarly to 4.11 and the standard arguments for approximation of functions in expectation, it is sufficient to prove (4.28) for the f such that f is differentiable and its derivative is bounded. Let $R_0(t) = \sup\{u \geq 0; N_0(t, t+u] = 0\}$. That is, $R_0(t)$ is the remaining time to count the next point of N_0 at time t. For nonnegative number s, let

$$Y(t) = f(X(t))e^{-sR_0(t)}, \qquad t \in \mathbb{R}.$$

Then, clearly $Y(t)$ is bounded and consistent with θ_t . Since $R_0'(t) = -1$, the right-hand derivative of $Y(t)$ is computed as

$$Y'(t) = (X'(t)f'(X(t)) + sf(X(0)))e^{-sR_0(t)}.$$

We next define a point process N_1 as

$$N_1(B) = N(B) - \max(N(B), N_0(B)), \qquad B \in \mathcal{B}(\mathbb{R}),$$

where it is noted that N is the point process given in the assumption (4.8d). Obviously, N_0 and N_1 are simple, and do not have a common point. Furthermore, N_1 is consistent with θ_t and has the intensity $\lambda_1 \equiv E(N_1(0,1]) < \lambda < \infty$. Thus, we can apply 4.5 (see (a) in Remark 4.3). Let $\varphi(s) = E(e^{-sR_0(0)})$, then

$$\begin{aligned} &E\left(X'(0)f'(X(0)) + sf(X(0))\right)\right)\varphi(s) \\ &\qquad = \lambda_0\left(E_0(f(X(0-))) - E_0(f(X(0)))\varphi(s)\right) \\ &\qquad\quad + \lambda_1 E_1\left(f(X(0-)) - f(X(0))\right)\varphi(s). \end{aligned} \tag{4.29}$$

By the memoryless property of the exponential distribution, we have $\varphi(s) = \lambda_0/(s+\lambda_0)$. Hence, letting $s \to \infty$ in (4.29) and using the fact that $s\varphi(s) \to \lambda_0$ and $\varphi(s) \to 0$, we obtain (4.28). □

Remark 4.4. If the reader is familiar with a martingale and the fact that the Poisson process N of Theorem 4.5 can be expressed as

$$N((0,t]) = \lambda t + M(t), \qquad t \geq 0,$$

where $M(t)$ is an integrable martingale with respect to the filtration $\sigma(X(u); u \leq t)$. See Section 4.11 for the definition of the martingale. Then, (4.28) is almost immediate from the definition of the Palm distribution since

$$\int_0^t f(X(u-))dM(u)$$

is also a martingale, and therefore its expectation vanishes. This proof is less elementary than the above proof.

Let us apply Theorem 4.5 together with the rate conservation law to the $M/G/1$ queue, which is a single server queue with the Poisson arrivals and independently and identically distributed requirements.

Example 4.7 (Pollaczek-Khinchine formula). We consider the special case of the workload process in Example 4.6. We here further assume that the processing rate $r(x) \equiv 1$, N is the Poisson process with rate $\lambda > 0$ and $\{S_n\}$ is a sequence of *i.i.d.* (independent and identically distributed) random variables which are independent of everything else.

Thus, we consider the workload process $V(t)$ of the $M/G/1$ queue. The service discipline of this queue can be arbitrary as long as the total service rate is always unit and the server can not be idle when there is a customer in the system. This process is known to be stable, that is, its stationary distribution exists if and only if

$$\rho \equiv \lambda E(S_1) < 1.$$

We assume this stability condition. Then, $\{V(t)\}$ is a stationary process under the stationary distribution. For nonnegative number θ, let $f(x) = e^{-\theta x}$ for $x \geq 0$, then (4.26) yields

$$\begin{aligned} \theta E(e^{-\theta V(0)} 1(V(0) > 0)) &= \lambda E_0(e^{-\theta V(0-)} - e^{-\theta(V(0-)+S_0)}) \\ &= \lambda E(e^{-\theta V(0-)})(1 - E(e^{-\theta S_1})), \end{aligned} \tag{4.30}$$

where we have used the *i.i.d.* assumption of S_n and Theorem 4.5 to get the second equality. We can rewrite the left-hand side as

$$\theta E(e^{-\theta V(t)}) - \theta P(V(0) = 0).$$

Let $\varphi(\theta) = E(e^{-\theta V(t)})$ and $g(\theta) = E(e^{-\theta S_1})$, then we have, from (4.30),

$$\theta\varphi(\theta) - \theta P(V(0) = 0) = \lambda\varphi(\theta)(1 - g(\theta)).$$

Since $\varphi(\theta) \to 1$ and $\frac{1-g(\theta)}{\theta} \to -g'(0) = E(S_1)$ as $\theta \downarrow 0$. we have $P(V(0) = 0) = 1 - \rho$, dividing the above formula by θ and letting $\theta \downarrow 0$. Thus, we have the Laplace-transform of $V(0)$ under the stationary assumption.

$$\varphi(\theta) = \frac{\theta(1-\rho)}{\theta - \lambda(1 - g(\theta))}, \qquad \theta > 0. \tag{4.31}$$

This formula is independently obtained by Pollaczek and Khinchine, and called Pollaczek-Khinchine formula. □

In this example, if the processing rate $r(x)$ is not a constant, then it is generally hard to get the stationary distribution of the workload in any form. We show that there is an exceptional case in the following example.

Example 4.8 (State dependent service). We again consider the workload process $V(t)$ in Example 4.6 under the assumptions of Example 4.7 except for the processing rate $r(x)$, which may be arbitrary. Thus, we consider the $M/G/1$ workload process with state dependent processing rate. In this case, the stability condition for $V(t)$ is complicated, so we here just assume that the stationary distribution exists. Similar to (4.30), we have

$$\theta E(r(V(0))e^{-\theta V(0)} 1(V(0) > 0)) = \lambda E(e^{-\theta V(0-)})(1 - E(e^{-\theta S_1})). \tag{4.32}$$

From this equation, it is generally hard to get $\varphi(\theta) \equiv E(e^{-\theta V(0-)})$ for a given $g(\theta) \equiv E(e^{-\theta S_1})$ except for the case that $r(x)$ is a constant.

We thus consider the special case that $r(x) = a + bx$ for nonnegative constant a and positive constant b. In this case, the left-hand side of (4.32) can be written as

$$\theta E((a + bV(0))e^{-\theta V(0)} 1(V(0) > 0)) = \theta(a\varphi(\theta) - b\varphi'(\theta)) - a\theta P(V(0) = 0).$$

Hence, we have the following differential equation from (4.32).

$$-\frac{1}{b}\left(a - \lambda\frac{1 - g(\theta)}{\theta}\right)\varphi(\theta) + \varphi'(\theta) = -\frac{a}{b}P(V(0) = 0) \tag{4.33}$$

For $\theta \geq 0$, let

$$h(\theta) = -\frac{1}{b}\int_0^\theta \left(a - \lambda\frac{1 - g(u)}{u}\right) du.$$

Then, the solution of (4.33) with the boundary condition $\varphi(0) = 1$ is obtained as

$$\varphi(\theta) = e^{-h(\theta)}\left(1 - \frac{a}{b}P(V(0) = 0)\int_0^\theta e^{h(u)} du\right).$$

To determine $P(V(0)=0)$, Note that $h(\infty) \equiv \lim_{\theta \to \infty} h(\theta) = -\infty$ if $a > 0$ while $h(\infty) = +\infty$ if $a = 0$. Hence, if $a = 0$, then

$$\varphi(\theta) = e^{-h(\theta)}.$$

If $a > 0$, then we must have $1 = \frac{a}{b} P(V(0)=0) \int_0^\infty e^{h(u)} du$, which concludes $P(V(0)=0) = \frac{b}{a}\left(\int_0^\infty e^{h(u)} du\right)^{-1}$. Hence, we finally have, for $a > 0$,

$$\varphi(\theta) = e^{-h(\theta)} \int_\theta^\infty e^{h(u)} du \left(\int_0^\infty e^{h(u)} du\right)^{-1}.$$

It may be interesting to see the mean workload $E(V(0)) = -\varphi'(0)$, which is

$$E(V(0)) = \begin{cases} \frac{1}{b}\rho, & a = 0, \\ \frac{1}{b}(\rho - a) + \left(\int_0^\infty e^{h(u)} du\right)^{-1}, & a > 0. \end{cases}$$

Note that these computations are not valid for $b = 0$. □

4.10 Relationship among the queueing length processes observed at different points in time

The rate conservation is powerful for complicated systems. This is exemplified for the system queue length process, i.e., the total number of customer in system, under a very general setting. Here, the queueing system is meant a service system with arrivals and departures. Let N_a and N_d be point processes composed of arrival and departure instants, respectively. We here allow those point processes to be not simple. Then, the the system queue length $L(t)$ at time t is defined as

$$L(t) = L(0) + N_a((0,t]) - N_d((0,t]), \qquad t \geq 0.$$

Note that customers who leave the system immediately after their arrivals without any service are counted as departure.

Theorem 4.6. For a queue system with arrival point process N_a, departure point process N_d and the system queue length $L(t)$, assume that P is θ_t-stationary and $N_a, N_d, L(t)$ are consistent with θ_t, that is, $(N_a, N_d, \{L(t)\})$ is jointly stationary. Let N_a^* and N_d^* be the simple versions of N_a and N_d. If $\lambda_a^* \equiv E(N_a^*((0,1])$ and $\lambda_d^* \equiv E(N_d^*((0,1]))$ are finite, then, for $n \in \mathbb{Z}_+$,

$$\lambda_a^* P_a^*(n+1-\triangle L(0) \leq L(0-) \leq n) = \underline{\lambda}_d^* \underline{P}_d^*(n+1+\triangle L(0) \leq L(0) \leq n), \tag{4.34}$$

where $\triangle L(0) = L(0) - L(0-)$, and P_a^* and $\underline{P}_d^*$ are Palm distributions of N_a^* and the following point process, respectively.

$$\underline{N}_d^*(B) = N_d^*(B) - \min(N_a^*(B), N_d^*(B)), \qquad B \in \mathcal{B}(\mathbb{R}),$$

and $\underline{\lambda}_d^*$ is its intensity.

Proof. We can apply 4.5 for $X(t) = 1(L(t) \geq n+1)$ and $N = N_a^* + \underline{N}_d^*$ because $X(t)$ is bounded and $X'(t) = 0$. Hence, (4.27) yields

$$\lambda_a^*(P_a^*(L(0-) \geq n+1) - P_a^*(L(0-) + \triangle L(0) \geq n+1))$$
$$+\underline{\lambda}_d^*(\underline{P}_d^*(L(0) - \triangle L(0) \geq n+1) - \underline{P}_d^*(L(0) \geq n+1)) = 0,$$

which concludes (4.34). □

In Theorem 4.6, $\underline{N}_d^*$ count instants when departures only occur.

Example 4.9 (Queueing model with no customer loss). In the model of Theorem 4.6, assume that there is no lost customer, and customers singly arrive and singly depart. Furthermore assume that arrivals and departures do not simultaneously occur. That is, $P_a(\triangle L(0) = 1) = P_d(\triangle L(0) = -1) = 1$, where P_d is the Palm distribution of N_d. From (4.34), it follows that

$$\lambda_a P_a(L(0-) = n) = \lambda_d P_d(L(0) = n), \qquad n = 0, 1, \ldots.$$

Summing both sides of the above equation over all n, we have $\lambda_a = \lambda_d$. Hence, we have

$$P_a(L(0-) = n) = P_d(L(0) = n), \qquad n \in \mathbb{Z}_+. \tag{4.35}$$

Thus, the system queue length observed by arriving customers is identical with the one observed by departing customers. This is intuitively clear, but it is also formally obtained, which is important to consider more complex situations. □

Example 4.10 (Loss system). For the queueing system of Example 4.9, assume that the system queue length is limited to M, and arriving customers who find M customers in system are lost. In this case, (4.35) does not hold generally. For example, its right-hand side vanishes for $n = M$, but its left-hand side may not be zero. Since both sides of (4.34) vanishes for $n \geq M$, we have

$$\lambda_a P_a(L(0-) = n) = \lambda_d P_d(L(0) = n), \qquad n = 0, 1, \ldots, M-1. \tag{4.36}$$

Since the departure reduces one customer, $P_d(L(0) \leq M-1) = 1$. Hence, summing (4.36) over for $n = 0, 1, \ldots, M-1$, we have

$$P_a(L(0-) = M) = \frac{\lambda_a - \lambda_d}{\lambda_a}.$$

This is the probability that customers are lost, and is again intuitively clear. We here correctly present it using Palm distributions. □

Theorem 4.6 does not have information on the system queue length at an arbitrary point in time. Let us include this information using supplementary variables.

Theorem 4.7. Under the assumptions of Theorem 4.6, let $R_a(t)$ be the time to the next arrival instant measure from time t, i.e., remaining arrival time, and let T_1 be the first arrival time after time 0. For a nonnegative measurable function f on $\mathbb{R}$ such that it is differentiable, $\overline{E}_a(f(T_1)) < \infty$ and $|E(f'(R_a(0)))| < \infty$, where $\overline{E}_a$ represents the expectation concerning the detailed Palm distribution of N_a, we have

$$\begin{aligned}
&-E(f'(R_a(0)); L(0) \geq n+1) \\
&\quad = \lambda_a \left(\overline{E}_a(f(0); L(0-) \geq n+1) - \overline{E}_a(f(T_1); L(0) \geq n+1)\right) \\
&\qquad + \underline{\lambda}_d \overline{\underline{E}}_d(f(R_a(0)); n+1+\triangle L(0) \leq L(0) \leq n), \qquad n \in \mathbb{Z}_+, \quad (4.37)
\end{aligned}$$

where $\overline{\underline{E}}$ represents the expectation of the detailed Palm distribution of $\underline{N}_d$, and λ_a and $\underline{\lambda}_d$ are the intensities of N_a and $\underline{N}_d$, respectively.

Proof. Let $X(t) = f(R_a(t))1(L(t) \geq n+1)$. Since $R_a'(t) = -1$, we have $X'(t) = -f'(R_a(t))1(L(t) \geq n+1)$. Since $P_a(R_a(0) = T_1) = 1$, 4.5 and the remark on the detailed version of the rate conservation law at the end of Section 4.8 concludes (4.37). □

Example 4.11 (NBUE distribution). In Example 4.9, assume that the interarrival times of customers are independent and identically distributed and that arrivals and departures do not simultaneously occur. Furthermore, assume that the interarrival time T_1 satisfies

$$E_a(T_1 - x | T_1 > y) \leq E_a(T_1), \qquad x \geq 0. \qquad (4.38)$$

The distribution of T_1 under P_a satisfying this condition is said to be NBUE type, where NBUE is the abbreviation of New Better than Used in Expectation. In fact, (4.38) represents that the conditional expectation of the remaining arrival time is not greater than the mean interarrival time. If the inequality in (4.38) is reversed, then the distribution of T_1 is said to be NWUE, which is the abbreviation of New Worse than Used in Expectation. Form the NBUE assumption, we have

$$E(R_a(0); L(0) \geq n+1) \leq E_a(T_1)P(L(0) \geq n+1).$$

We apply Theorem 4.7 with $f(x) = x$. Since $f(0) = 0$, $f'(x) = 1$ and $\lambda_a = \underline{\lambda}_d = \lambda_d = 1/E_a(T_1)$, (4.37) yields

$$-P(L(0) \geq n+1) \leq -\overline{P}_a(L(0) \geq n+1) + \overline{\underline{P}}_d(L(0) = n), \qquad n \geq 0.$$

Since $P_a = \overline{P}_a$ and $P_d = \overline{\underline{P}}_d$, this and (4.35) lead to

$$P_a(L(0-) \geq n+1) \leq P(L(0) \geq n+1), \qquad n \geq 0. \qquad (4.39)$$

Hence, the distribution of the system queue length at the arrival instants is greater than the one at an arbitrary point in time in stochastic order, where, for two distribution functions F and G, F is said to be greater than G in stochastic order if $1-F(G) \geq 1-G(x)$ for all $x \in \mathbb{R}$. □

Similarly to Theorem 4.7, we can take the minimum of the remaining service times of customers being served, and get relationships among the distributions of the system queue lengths at different embedded points in time.

4.11 An extension of the rate conservation law

In this section, we briefly discuss how the rate conservation law (4.23) can be generalized for other types of processes. For this, it is notable that this law is obtained from the integral representation of the time evolution (4.24) and the definition of Palm distribution P_0. There are two integrators, du of the Lebesgue measure and $N(du)$ of a point process, both of which are defined on the line. To closely look at this, we rewrite (4.24) in a slightly extended form as

$$X(t) = X(0) + \int_0^t X'(u)A(du) + \int_0^t \Delta X(u)N(du),$$

where $A(t) - A(0)$ is consistent with the shift operator θ_t and has bounded variations, and $\Delta X(u) = X(u) - X(u-)$. If $X(t)$ has either a component of unbounded variations or a continuous and singular component with respect to the Lebesgue measure, this expression breaks down. To get back the expression, we subtract this component, denoting it by $M(t)$. Thus, we have

$$X(t) - M(t) = X(0) - M(0) + \int_0^t Y'(u)A(du) + \int_0^t \Delta Y(u)N(du),$$

where $Y(u) = X(u) - M(u)$. If $M(t)$ is consistent with θ_t, then we have the rate conservation law for the process $\{Y(t)\}$. It may be reasonable to assume that $M(t)$ is continuous. However, this rate conservation law may not be useful to study $\{X(t)\}$ because $X(t)$ is not directly involved.

To get useful information, we make use of a test function, which is used in 4.4, and apply Itô's integration formula, assuming that $M(t)$ is a square integrable martingale. That is,

(4.11a) $M(t)$ is continuous in t and consistent with $\{\theta_t\}$.
(4.11b) $E((M(t) - M(0))^2) < \infty$ for all $t \geq 0$.
(4.11c) $\{M(t) - M(0); t \geq 0\}$ is a martingale with respect to $\{\mathcal{F}_t\}$, that is,

$$E(M(t) - M(0)|\mathcal{F}_s) = M(s) - M(0), \qquad 0 \leq s \leq t,$$

where $\mathcal{F}_t$ is a sub σ-field of $\mathcal{F}$ which is increasing in $t \in \mathbb{R}$, and $\{\mathcal{F}_t; t \in \mathbb{R}\}$ is called a filtration.

This martingale assumption is typical for a process with unbounded variations. It is beyond our scope to fully discuss Itô's integration formula, but we like to see how it works. The reader may refer to standard text books such as [16] and [17] for more details. Assume that $X(t)$ and $M(t)$ are $\mathcal{F}_t$-measurable for all $t \in \mathbb{R}$.

For convenience, let $M_0(t) = M(t) - M(0)$ for $t \geq 0$. Under these assumptions, $M_0^2(t)$ is submartingale, that is,

$$E(M_0^2(t)|\mathcal{F}_s) \geq M_0^2(s), \qquad 0 \leq s \leq t,$$

and there exists a nondecreasing process $\langle M_0(t)\rangle$ such that $M_0^2(t) - \langle M_0(t)\rangle$ is a martingale. Then, Itô's integration formula reads: for twice continuously differentiable function f,

$$\begin{aligned} f(X(t)) = f(X(0)) &+ \int_0^t f'(X(u))dM(u) + \int_0^t f'(X(u))Y'(u)A(du) \\ &+ \frac{1}{2}\int_0^t f''(X(u))d\langle M_0(u)\rangle + \int_0^t \Delta f(Y(u))N(du), \quad (4.40) \end{aligned}$$

where the integration on the interval $[0,t]$ with respect to $dM(u)$ is defined L^2-limit of the Riemann sum, that is, $\sum_{\ell=1}^n f'(X(\frac{\ell-1}{n}))(M(\frac{\ell}{n}) - M(\frac{\ell-1}{n}))$. See Theorems 17.18 and 26.6 of [16] and Theorem 3.3 of [17]. This integration is a martingale, and its expectation vanishes. Define the Palm distribution with respect to $\langle M_0(t)\rangle$ as

$$P_{\langle M\rangle}(C) = \frac{1}{\lambda_{\langle M\rangle}} E\left(\int_0^1 1_C \circ \theta_u d\langle M_0(u)\rangle\right), \qquad C \in \mathcal{F}.$$

where $\lambda_{\langle M\rangle} = E(M(1) - M(0))$. The Palm measure P_A is similarly defined for the non-decreasing process A. Thus, taking the expectation of both sides of (4.40), we arrive at

$$E_A(f'(X(0))Y'(0)) + \frac{1}{2}\lambda_{\langle M\rangle}E_{\langle M\rangle}(f''(X(0))) + \lambda E_0(\Delta f(Y(0))) = 0, \quad (4.41)$$

assuming suitable finiteness conditions for the expectations, where E_A and $E_{\langle M\rangle}$ stand for the expectations concerning P_A and $P_{\langle M\rangle}$.

We can proceed one further step using the representation theorem for a continuous martingale by the Brownian motion. This theorem says that, for a continuous martingale $M_0(t)$ with respect filtration $\{\mathcal{F}_t\}$, there exists a progressively measurable process $Z(t)$ such that

$$M_0(t) = \int_0^t Z(u)dB(u), \qquad \langle M_0(t)\rangle = \int_0^t Z^2(u)du < \infty, \qquad t \geq 0,$$

where $\{Z(t)\}$ is said to be progressively measurable if $\{(u,\omega) \in [0,t] \times \Omega; Z(u) \in A\} \in \mathcal{B}([0,t]) \times \mathcal{F}_t$ for all $t \geq 0$, and $\{B(t); t \geq 0\}$ is called a Brownian motion if it has independent and stationary increments which are normally distributed with mean 0 and unit variance (see, e.g., Theorem 18.12 of [16] and Theorem 4.15 of

[17]). Hence, (4.41) can be written as

$$E_A(f'(X(0))Y'(0)) + \frac{\lambda_{Z^2}}{2} E\left(f''(X(0))Z^2(0)\right) + \lambda E_0(\Delta f(Y(0))) = 0, \quad (4.42)$$

where $\lambda_{Z^2} = E(\int_0^1 Z^2(u)du)$.

In queueing applications of (4.42), $Y(t)$ and $Z(t)$ are often identified as functions of $X(t)$ from their modeling assumptions through the expression:

$$X(t) = X(0) + \int_0^t Y'(u)A(du) + \int_0^t Z(u)dB(u) + \int_0^t \Delta Y(u)N(du). \quad (4.43)$$

In this case, $Y(t) = g(X(t))$ and $Z(t) = h(X(t))$ for some functions g and h, and (4.42) is really useful to consider the stationary distribution of $X(t)$.

Example 4.12 (extended Pollaczek-Khinchine formula). Let us consider to add the Brownian motion to the workload process $V(t)$ in Example 4.7. That is, $V(t)$ is changed to the following $X(t)$.

$$\begin{aligned} X(t) &= X(0) + \sigma^2 B(t) + \sum_{n=1}^{N(t)} S_n - t + I(t) \\ &= X(0) + \int_0^t (I(du) - du) + \int_0^t \sigma^2 dB(u) + \int_0^t \Delta Y(u)N(du) \end{aligned}$$

where $I(t)$ is a minimum non-decreasing process for $X(t)$ to be nonnegative. That is, $I(t)$ is a regulator. Thus, if we put $A(t) = I(t) - t$, $Y(t) = t + \int_0^t \Delta Y(u)N(du)$ and $Z(t) = \sigma$, then we have (4.43).

Assume the stability condition that

$$\rho \equiv \lambda E(S_1) < 1.$$

Then, $\{X(t)\}$ is a stationary process under the stationary distribution. For nonnegative number θ, let $f(x) = e^{-\theta x}$. We apply (4.42) to $X(t)$ and this f. Since $I(t)$ is increased only when $X(t) = 0$, we have, using $\varphi(\theta) = E(e^{-\theta X(0)})$ and $g(\theta) = E(e^{-\theta S_1})$,

$$\theta(\varphi(\theta) - E_I(1)) + \frac{\sigma^2\theta^2}{2}\varphi(\theta) = \lambda\varphi(\theta)(1 - g(\theta)).$$

Similar to Example 4.7, we have $E_I(1) = 1 - \rho$, dividing the above formula by θ and letting $\theta \downarrow 0$. Thus, we have the Laplace-transform of $X(0)$ under the stationary assumption.

$$\varphi(\theta) = \frac{\theta(1-\rho)}{\theta + \frac{1}{2}\sigma^2\theta^2 - \lambda(1 - g(\theta))}, \qquad \theta > 0. \quad (4.44)$$

This is an extension of the Pollaczek-Khinchine formula (4.31). □

The results of the present section can be obtained under weaker assumptions and for a multidimensional process. The latter is in a similar line to 4.4 with a multidimensional version of the Itô integration formula, while the continuous martingale can be weakened to a local martingale with unbounded variational discontinuity. Of course, we need to carefully consider the integration under such discontinuity.

4.12 Piece-wise deterministic Markov process (PDMP)

As we discussed in Section 4.1, many queueing models can be described by stochastic processes whose major changes occur in embedded points in time. In this section, we introduce a typical Markov process having such structure. The sample path of this Markov process is assumed to satisfy the integral representation (4.24) and to have discontinuous points only on a set, called boundary. It will be shown that this process is flexible and has a wide range of applications.

We first introduce notation for state spaces. Let $\mathcal{X}$ be a countable set. An element $\mathbf{x} \in \mathcal{X}$ is referred to as a macro state. For each $\mathbf{x} \in \mathcal{X}$, let $K_{\mathbf{x}}$ be a closed subset of $\mathbb{R}^{m(\mathbf{x})}$, where $m(\mathbf{x})$ be a positive integer determined by $\mathbf{x}$ and $\mathbb{R}^n$ is the n-dimensional Euclid space, i.e., vector space with the Euclidean metric. Define sets K and $J(\mathbf{x})$ as

$$K = \{(\mathbf{x},\mathbf{y}); \mathbf{x} \in \mathcal{X}, \mathbf{y} \in K_{\mathbf{x}}\}, \qquad J(\mathbf{x}) = \{1,2,\dots,m(\mathbf{x})\}.$$

For $(\mathbf{x},\mathbf{y}) \in K$, $\mathbf{y}$ is referred to as a continuous component or supplementary variable under macro state $\mathbf{x}$.

On this K, we introduce a natural topology induced from those on $K_{\mathbf{x}}$. For each $\mathbf{z} \equiv (\mathbf{x},\mathbf{y}) \in K$, the family of its neighborhoods is generated by all the sets of the form $\{\mathbf{x}\} \times (V_{\mathbf{y}} \cap K_{\mathbf{x}})$, where $V_{\mathbf{y}}$ is a neighborhood of $\mathbf{y} \in \mathbb{R}^{m(\mathbf{x})}$. Let $\mathcal{B}(K)$ be the Borel σ-field on K, i.e., the σ-field generated by all open sets of K. Thus, $(K,\mathcal{B}(K))$ is measurable space and we can define a probability measure on it.

We further need notation on boundary. Let $K_{\mathbf{x}+}$ be an open subset of $K_{\mathbf{x}}$, and let $K_{\mathbf{x}0} \equiv K_{\mathbf{x}} \setminus K_{\mathbf{x}+}$, which is called a boundary. For K, we define its inside K_+ and its boundary K_0 as

$$K_+ = \{(\mathbf{x},\mathbf{y}); \mathbf{x} \in \mathcal{X}, \mathbf{y} \in K_{\mathbf{x}+}\}, \qquad K_0 = K \setminus K_+.$$

Definition 4.10 (PDMP). Let $\mathbf{Z}(t) \equiv (\mathbf{X}(t),\mathbf{Y}(t))$ be a stochastic process with state space K defined above, and assume that $\mathbf{Z}(t)$ is right-continuous with left-limits. This $\{\mathbf{Z}(t)\}$ is said to be a piece-wise deterministic Markov process, PDMP for short, if the following three conditions are satisfied.

(4.12a) $\mathbf{X}(t)$ is unchanged as long as $\mathbf{Y}(t) \equiv (Y_1(t),\dots,Y_{m(\mathbf{x})}(t)) \in K_{\mathbf{x}+}$, which changes according to the following differential equation when $\mathbf{X}(t) = \mathbf{x}$.

$$\frac{dY_\ell(t)}{dt} = g_{\mathbf{x}\ell}(\mathbf{Y}(t)), \qquad \ell \in J(\mathbf{x}),$$

where $g_{\mathbf{x}\ell}$ is a bounded measurable function from $\mathbb{R}^{m(\mathbf{x})}$ to $\mathbb{R}$ for each $\mathbf{x} \in \mathcal{X}$, and $\mathbf{Y}(t)$ hits boundary $K_{\mathbf{x}0}$ in a finite time with probability one. We refer to $\mathbf{X}(t)$ and $\mathbf{Y}(t)$ as macro state and continuous component, respectively.

(4.12b) At the moment when $\mathbf{Z}(t)$ hits the boundary K_0, that is, $\mathbf{Z}(t-) \in K_0$, it instantaneously returns to the inside, that is, $\mathbf{Z}(t-) \in K_0$ is changed to $\mathbf{Z}(t) \in K_+$ subject to the transition kernel Q from the boundary K_0 to the inside K_+. That is, for each $(\mathbf{x},\mathbf{y}) \in K_0$,

$$P(\mathbf{Z}(t) \in A | \mathbf{Z}(t-) = (\mathbf{x},\mathbf{y})) = Q((\mathbf{x},\mathbf{y}),A), \qquad A \in \mathcal{X} \times \mathcal{B}(K).$$

Q is referred to as a jump transition kernel.

(4.12c) For each finite time interval, the number of the hitting times at the boundary, i.e., the number of t such that $Y(t-) \in K_0$ is finite. We denote the point process generated by such hitting times by N.

Remark 4.5. The PDMP was introduced by Davis [10] (see also [11]). However, our definition of PDMP is slightly different from his definition. They use the attained lifetimes for the supplementary variables $\mathbf{Y}(t)$. Thus, the macro state transitions randomly occur subject to intensity depending on $\mathbf{Y}(t)$, which may hit the boundary. However, if the time is reversed, then their PDMP becomes ours. A minor advantage of ours is that the existence of the intensity is not necessary. This means that we do not need to assume the existence of densities of lifetime distributions for macro state transitions, which will be discussed below. □

The PDMP (piece-wise deterministic Markov process) looks complicated, but it has simple structure when we only observe the embedded epochs due to the state transition by Q. Let $\{t_n; n \in \mathbb{Z}\}$ be the set of such epochs numbered in increasing order. Then, $\{\mathbf{Z}(t_n-)\}$ is a discrete time embedded Markov chain. Let us consider the transition kernel of this embedded Markov chain.

For each state $\mathbf{z} \equiv (\mathbf{x},\mathbf{y}) \in K_+$, denote the time to the next transition starting from this state by $\zeta(\mathbf{x},\mathbf{y})$, which is uniquely determined by (4.12a). We let $\zeta(\mathbf{x},\mathbf{y}) = 0$ if $(\mathbf{x},\mathbf{y}) \in K_0$. We also denote the state of the continuous component $\mathbf{Y}(t)$ that attains just before this time by $\psi(\mathbf{x},\mathbf{y})$. Let H be the transition kernel H of the embedded process $\{\mathbf{Z}(t_n-)\}$. That is,

$$H(\mathbf{z},\{\mathbf{x}'\} \times B) = P(\mathbf{Z}(t_{n+1}-) \in \{\mathbf{x}'\} \times B | \mathbf{Z}(t_n-) = \mathbf{z}),$$
$$\mathbf{z} \in K_0, \mathbf{x}' \in \mathcal{X}, B \in \mathcal{B}(K_{\mathbf{x}'}).$$

Then, it is easy to see that

$$H(\mathbf{z},\{\mathbf{x}'\} \times B) = \int_{K_{\mathbf{x}'}} Q(\mathbf{z},\{\mathbf{x}'\} \times d\mathbf{y}') 1(\psi(\mathbf{x}',\mathbf{y}') \in B). \tag{4.45}$$

Example 4.13. As an example of PDMP, let us consider the workload process $V(t)$ of Example 4.8 for the $M/G/1$ queue with state dependent processing rate r. Let $X(t) \equiv 0$, $Y_1(t) = t - T_{N(t)}$ and $Y_2(t) = V(t)$, then $Y_1'(t) = -1$ and $Y_2'(t) = r(V(t))1(V(t) > 0)$. Hence, if we let $\mathcal{X} = \{0\}$ and $K = \{0\} \times [0,\infty)^2$ with $K_0 =$

$\{0\}^2 \times [0,\infty)$, then $(X(t),(Y_1(t),Y_2(t)))$ is a PDMP, where the jump transition Q is given by

$$Qf(0,(0,x)) = E(f(0,(T_1,x+S_1))), \qquad x \geq 0,$$

for a nonnegative valued function f on $K_+ \equiv \{0\} \times (0,\infty) \times [0,\infty)$. Note that $Y_2(t)$ has no boundary in this formulation. □

We next to consider the stationary distribution of PDMP (piece-wise deterministic Markov process). We are interested to characterize it using the rate conservation law. We first consider its transition operator of the Markov process $\{\mathbf{Z}(t)\}$. Since its state space K includes continuous components, we consider the transition operator to work on the space of suitable functions on K.

Let $\mathcal{M}_b(K)$ be the set of all bounded functions from $\mathcal{K}$ to $\mathbb{R}$ which are $\mathcal{B}(K)/\mathcal{B}(\mathbb{R})$-measurable. For each $t \geq 0$, define operator T_t on $\mathcal{M}_b(K)$ as

$$T_t f(\mathbf{z}) = E(f(\mathbf{Z}(t))|\mathbf{Z}(0) = \mathbf{z}), \qquad \mathbf{z} \in K, f \in \mathcal{M}_b(K).$$

Note that T_t is a linear function from $\mathcal{M}_b(K)$ to $\mathcal{M}_b(K)$. Furthermore, it maps a nonnegative function to a nonnegative function. Thus, T_t is nonnegative and linear operator on $\mathcal{M}_b(K)$, which uniquely determines a distribution on $(K,\mathcal{B}(K))$ as is well known.

Define operator $\mathcal{A}_+$ as

$$\mathcal{A}_+ f(\mathbf{z}) = \lim_{t\downarrow 0} \frac{1}{t}(T_t f(\mathbf{z}) - f(\mathbf{z})), \qquad \mathbf{z} \in K_+,$$

as long as it exists. We refer to this $\mathcal{A}_+$ as a weak generator. Let $\mathcal{D}_{\mathcal{A}_+}$ be the set of all $f \in \mathcal{M}_b(K)$ such that $\mathcal{A}_+ f$ exists. Note that $\mathcal{A}_+$ is a generator only for the continuous part of $\mathbf{Z}(t)$, and does not include the information on state changes due to the macro state transitions. Hence, $\mathcal{A}_+$ is not a generator in the sense that it determines the operator T_t. This is the reason why we call it weak.

For each macro state $\mathbf{x} \in \mathcal{X}$, let $\mathcal{Y}_\mathbf{x}$ be the set of all solutions $\{\mathbf{y}(t)\}$ for the differential equation (4.12a), i.e.,

$$\frac{dy_\ell(t)}{dt} = g_\mathbf{x}(\mathbf{y}(t)), \qquad 0 \leq t < \zeta(\mathbf{x},\mathbf{y}).$$

Let $\mathcal{M}_b^1(K)$ be the set of all functions $f \in \mathcal{M}_b(K)$ such that $f(\mathbf{x},\xi(t))$ has the right-hand derivative in all t in the domain of ξ and is continuous from the left at $t = \zeta(\mathbf{x},\mathbf{y})$ for $\mathbf{x} \in \mathcal{X}$ and $\xi \in \mathcal{Y}_\mathbf{x}$. Let $C_b^1(K)$ be the set of all functions $f \in \mathcal{M}_b(K)$ such that $f(\mathbf{x},\mathbf{y})$ has bounded and continuous partial derivatives $\frac{\partial}{\partial y_\ell} f(\mathbf{x},\mathbf{y})$ $(\ell = 1,2,\ldots,m(\mathbf{x}))$ for each $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in K_{\mathbf{x}+}$. Clearly, $C_b^1(K) \subset \mathcal{M}_b^1(K)$.

For $f \in \mathcal{M}_b^1(K)$ and $\mathbf{Z}(t) \in K$, it follows from the definition of the PDMP that

$$f(\mathbf{Z}(t)) - f(\mathbf{Z}(0)) = \int_0^t \frac{d}{du} f(\mathbf{Z}(u))du + \int_0^t (f(\mathbf{Z}(u)) - f(\mathbf{Z}(u-)))N(du).$$

Hence, for $\mathbf{z} \equiv (\mathbf{x}, \mathbf{y}) \in K_+$ and $\xi \in \mathcal{M}_b^1(K)$ with $\xi(0) = \mathbf{y}$, we have

$$\begin{aligned} \mathcal{A}_+ f(\mathbf{z}) &= \lim_{t \downarrow 0} \frac{1}{t} E\left(f(\mathbf{Z}(t)) - f(\mathbf{Z}(0)) | \, \mathbf{Z}(0) = \mathbf{z} \right) \\ &= \lim_{t \downarrow 0} \frac{1}{t} E\left(\int_0^t \frac{d}{du} f(\mathbf{Z}(u)) du \right| \mathbf{Z}(0) = \mathbf{z} \Bigg) \\ &= \left. \frac{d}{du} f(\mathbf{x}, \xi(u)) \right|_{u=0}, \end{aligned}$$

where the second equality is obtained since $\mathbf{Z}(u)$ must stay in K_+ for a finite time under the condition that $\mathbf{Z}(0) = \mathbf{z} \in K_+$. In particular, for $f \in C_b^1(K)$,

$$\mathcal{A}_+ f(\mathbf{z}) = \sum_{\ell=1}^{m(\mathbf{x})} g_{\mathbf{x}\ell}(\mathbf{y}) \frac{\partial}{\partial y_\ell} f(\mathbf{x}, \mathbf{y}). \tag{4.46}$$

Hence, $C_b^1(K) \subset \mathcal{M}_b^1(K) \subset \mathcal{D}_{\mathcal{A}_+}$. However, $C_b^1(K) \neq \mathcal{M}_b^1(K)$ in general. For example, $\zeta \in \mathcal{M}_b^1(K)$, but $\zeta \notin C_b^1(K)$ after Definition 4.11.

Theorem 4.8. Let $\{\mathbf{Z}(t)\}$ be the PDMP and let N be the point process N generated by hitting times at the boundary. If $\{\mathbf{Z}(t)\}$ has the stationary distribution ν and if N has a finite intensity λ, then there exists a probability distribution ν_0 on $(K_0, \mathcal{B}(K_0))$ satisfying

$$\int_{K_+} \mathcal{A}_+ f(\mathbf{z}) \nu(d\mathbf{z}) = \lambda \int_{K_0} (f(\mathbf{z}) - Qf(\mathbf{z})) \nu_0(d\mathbf{z}), \qquad f \in \mathcal{M}_b^1(K). \tag{4.47}$$

Conversely, if there exist probability distributions ν on $(K_+, \mathcal{B}(K_+))$ and ν_0 on $(K_0, \mathcal{B}(K_0))$ satisfying (4.47) with some positive number λ, then

$$\overline{\nu}(B) = \nu(B \cap K_+), \qquad B \in \mathcal{B}(K)$$

is the stationary distribution of $\mathbf{Z}(t)$, and the point process N has the finite intensity λ. Furthermore, let P be a probability measure on $(\Omega, \mathcal{F})$ such that $\{\mathbf{Z}(t)\}$ is the stationary process with the stationary distribution ν, then ν_0 is the distribution of $\mathbf{Z}(0-)$ under the Palm distribution P_0 with respect to N.

Remark 4.6. Davis [11] computes an extended generator, which characterizes the stationary distribution, for the PDMP supplemented by the attained lifetimes. We can rewrite (4.47) in a similar form. Namely, let $\lambda(\mathbf{z}) = \lambda \frac{\nu_0(d\mathbf{z})}{\nu(d\mathbf{z})}$, where $\frac{\nu_0(d\mathbf{z})}{\nu(d\mathbf{z})}$ is the Radon Nikodym derivative of ν_0 with respect to ν. Then, we have

$$\int_K (\mathcal{A}_+ f(\mathbf{z}) + \lambda(\mathbf{z})(Qf(\mathbf{z}) - f(\mathbf{z}))) \, \nu(d\mathbf{z}) = 0. \tag{4.48}$$

$\lambda(\mathbf{z})$ can be considered as a stochastic intensity, and the integrant corresponds with the extended generator. (4.48) is particularly useful when $\lambda(\mathbf{z})$ is available, but this may not be always the case. In this situation, (4.47) is more flexible. □

Proof. Assume that $\{\mathbf{Z}(t)\}$ is a stationary process under probability measure P. Denote the stationary distribution of $\mathbf{Z}(t)$ by ν. Since the set of the times when $\mathbf{Z}(t)$ is on the boundary is countable, $P(\mathbf{Z}(0) \in K_0) = 0$. Hence, ν can be viewed as a probability distribution on $(K_+, \mathcal{B}(K_+))$. Let P_0 be the Palm distribution of P with respect to N. Since the distribution $\mathbf{Z}(0-)$ under P_0 is determined by ν_0, (4.47) is immediate from (4.46) and 4.4.

We next prove the converse. Suppose that there exists probability measures ν, ν_0 satisfying (4.47) and positive constant λ. Let $f \in \mathcal{M}_b(K)$. Since $T_u f$ is continuous in u, we have, from the definition of $\mathcal{A}_+$

$$\begin{aligned}\mathcal{A}_+ \left(\int_0^t T_u f du \right)(\mathbf{z}) &= \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \left(T_\varepsilon \left(\int_0^t T_u f du \right)(\mathbf{z}) - \int_0^t T_u f(\mathbf{z}) du \right) \\ &= \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \left(\int_0^t T_{u+\varepsilon} f(\mathbf{z}) du - \int_0^t T_u f(\mathbf{z}) du \right) \\ &= \lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} \left(\int_t^{t+\varepsilon} T_u f(\mathbf{z}) du - \int_0^\varepsilon T_u f(\mathbf{z}) du \right) \\ &= T_t f(\mathbf{z}) - f(\mathbf{z}), \qquad \mathbf{z} \in K_+. \end{aligned} \tag{4.49}$$

Define h for $f \in \mathcal{M}_b(K)$ as

$$h(\mathbf{z}) = \int_0^t T_u f(\mathbf{z}) du, \qquad \mathbf{z} \in K.$$

Then, (4.49) implies that $h \in \mathcal{D}_{\mathcal{A}_+}$. In general, h may not be in $\mathcal{M}_b^1(K)$, but we can prove that (4.47) holds for this h in the place of f by approximating h by functions in $\mathcal{M}_b^1(K)$. Since this proof is complicated, we omit it, but the reader can find it in [27]. Since, for $\mathbf{Z}(0-) \in K_0$,

$$\begin{aligned} QT_u f(\mathbf{Z}(0-)) &= E(T_u f(\mathbf{Z}(0)) | \mathbf{Z}(0-)) \\ &= E(f(\mathbf{Z}(u)) | \mathbf{Z}(0-)) = T_u f(\mathbf{Z}(0-)), \qquad u > 0, \end{aligned}$$

implies

$$\int_{K_0} T_u f(\mathbf{z}) \nu_0(d\mathbf{z}) = \int_{K_0} QT_u f(\mathbf{z}) \nu_0(d\mathbf{z}).$$

Integrating both sides for $u \in [0,t]$, we have

$$\int_{K_0} h(\mathbf{z}) \nu_0(d\mathbf{z}) = \int_{K_0} Qh(\mathbf{z}) \nu_0(d\mathbf{z}).$$

Hence substituting h into f of (4.47), (4.49) yields

$$\int_{K_+} T_t f(\mathbf{z}) \nu(d\mathbf{z}) = \int_{K_+} f(\mathbf{z}) \nu(d\mathbf{z}), \qquad t > 0.$$

Thus, $\overline{\nu}$ is the stationary distribution of $\mathbf{Z}(t)$.

We next prove that N has the finite intensity λ. To this end, define φ_ε for $\varepsilon > 0$ as

$$\varphi_\varepsilon(u) = \frac{1}{\varepsilon}\min(\varepsilon, u), \qquad u \geq 0.$$

We remind that $\zeta(\mathbf{x},\mathbf{y})$ is the hitting time at the boundary starting from the state $(\mathbf{x},\mathbf{y}) \in K$, where $\zeta(\mathbf{x},\mathbf{y}) = 0$ for $(\mathbf{x},\mathbf{y}) \in K_0$. For the trajectory $\xi \in \mathcal{Y}_\mathbf{x}$, $\frac{d}{dt}\zeta(\mathbf{x},\xi(t)) = -1$. Hence,

$$\frac{d}{dt}\varphi_\varepsilon(\zeta(\mathbf{x},\xi(t))) = -\frac{1}{\varepsilon}1(0 < \zeta(\mathbf{x},\xi(t)) \leq \varepsilon).$$

Let $f(\mathbf{x},\mathbf{y}) = \varphi_\varepsilon(\zeta(\mathbf{x},\mathbf{y}))$. Then, $f \in \mathcal{M}_b^1(K)$. We apply this f in (4.47), and let $\varepsilon \downarrow 0$. Then

$$\lim_{\varepsilon\downarrow 0}\int_{K_0}\varphi_\varepsilon(\zeta(\mathbf{x},\mathbf{y}))\nu_0(d\mathbf{x},d\mathbf{y}) = \int_{K_0}1(\zeta(\mathbf{x},\mathbf{y}) > 0)\nu_0(d\mathbf{x},d\mathbf{y}) = 0,$$

$$\lim_{\varepsilon\downarrow 0}\int_{K_0}\sum_{\mathbf{x}'\in\mathcal{X}} Q((\mathbf{x}',\mathbf{y}'),(\mathbf{x},\mathbf{y}))\varphi_\varepsilon(\zeta(\mathbf{x},\mathbf{y}))\nu_0(d\mathbf{x}',d\mathbf{y}') = \nu_0(K_0) = 1.$$

Here, we have used the fact that ν_0 is the distribution of $\mathbf{Z}(t)$ just before hitting the boundary K_0. Since ν is the stationary distribution, the above computations yield

$$\lim_{\varepsilon\downarrow 0}\frac{1}{\varepsilon}\int_{K_+}1\,(0 < \zeta(\mathbf{z}) \leq \varepsilon)\,\nu(d\mathbf{z}) = \lambda.$$

Reminding that N counts the hitting times at the boundary, we have

$$E(N((0,1]) = \lambda.$$

That is, λ is the intensity of N. Since λ is finite, we can define Palm distribution P_0 of P with respect to N. Denote the distribution of $\mathbf{Z}(0)$ under this P_0 by $\tilde{\nu}_0$, then the rate conservation law (4.23) yields

$$\int_{K_+}\mathcal{A}_+f(\mathbf{z})\nu(d\mathbf{z}) = \lambda\int_{K_0}(f(\mathbf{z}) - Qf(\mathbf{z}))\tilde{\nu}_0(d\mathbf{z}), \qquad f \in \mathcal{M}_b^1(K).$$

This together with (4.47) concludes

$$\int_{K_0}(f(\mathbf{z}) - Qf(\mathbf{z}))\tilde{\nu}_0(d\mathbf{z}) = \int_{K_0}(f(\mathbf{z}) - Qf(\mathbf{z}))\nu_0(d\mathbf{z}), \qquad f \in \mathcal{M}_b^1(K).$$

Here, we choose f_θ for f such that, for each $\mathbf{x} \in \mathcal{X}$ and $\theta_\ell \geq 0$ ($\ell = 1,2,\ldots,J(\mathbf{x})$),

$$f_\theta(\mathbf{x}',\mathbf{y}') = 1(\mathbf{x}' = \mathbf{x})\prod_{\ell\in J(\mathbf{x})} e^{-\theta_\ell y'_\ell}, \qquad (\mathbf{x}',\mathbf{y}') \in K_+.$$

For each subset U of $J(\mathbf{x})$, we let $\theta_\ell \to \infty$ for all $\ell \in U$. Then, we have $\nu_0 = \tilde{\nu}_0$ on the boundary $\{\mathbf{y} \in K_\mathbf{x}; y_\ell = 0$ for all $\ell \in U\}$ since $Qf_\theta(\mathbf{z})$ goes to zero. Changing U over all subsets of $J(\mathbf{x})$, we obtain $\tilde{\nu}_0 = \nu_0$ on K_0. This completes the proof. □

For applications, Theorem 4.8 is not convenient since $\mathcal{M}_b^1(K)$ is too large for verifying (4.47). We can replace it by a smaller class of functions.

Corollary 4.6. In Theorem 4.8, the condition (4.47) can be replaced by

$$\int_{K_{\mathbf{x}+}} \sum_{\ell=1}^{m(\mathbf{x})} g_{\mathbf{x}\ell}(\mathbf{y}) \frac{\partial}{\partial y_\ell} f(\mathbf{x},\mathbf{y}) \nu(d\mathbf{x} \times d\mathbf{y}) \\ = \lambda \int_{K_0} (f(\mathbf{z}) - Qf(\mathbf{z}))\nu_0(d\mathbf{z}), \qquad f \in C_b^1(K). \quad (4.50)$$

The necessity of (4.50) is immediate, but its sufficiency needs approximation arguments for functions in $\mathcal{M}_b^1(K)$ by those in $C_b^1(K)$. This argument can be found in [26], and omitted here.

4.13 Exponentially distributed lifetime

For the PDMP, some of its continuous components $Y_i(t)$ may be exponentially distributed and be decreased with constant rates. For example, this is the case when customers arrive subject to a Poisson process, and $Y_i(t)$ is the remaining time to the next arrival. In such a case, we can remove those continuous components to have a stochastically equivalent Markov process because the exponential distribution has memoryless property, that is, if random variable T has the exponential distribution, then

$$P(T > s+t | T > s) = P(T > t), \qquad s,t \geq 0.$$

Since this case is particularly interested in our applications, we make the following assumption.

(4.13a) The jump transition kernel Q of the PDMP does not depend on the continuous components which are exponentially distributed and decreased with constant rates.

In this section, we characterize the stationary distribution for this type of piece-wise deterministic Markov processes (PDMP).

Although we do not need to keep track such continuous components, the standard description of PDMP must include them. Thus, after removing them from the PDMP, we have to care about the modified process, which is not exactly PDMP. In this section, we are particularly interested in the stationary distribution of this modified process.

Assume that the PDMP satisfies the assumption (4.13a). For macro state $\mathbf{x} \in \mathcal{X}$ and continuous component $\mathbf{y} \in K_{\mathbf{x}}$, let $J_e(\mathbf{x})$ be the index set of $\mathbf{y}$'s which have exponential distributions. We denote the decreasing rate of the i-th component for $i \in J_e(\mathbf{x})$ by $c_{\mathbf{x}i} \geq 0$. We replace the i-th entry of $\mathbf{y}$ by 0 for $i \in J_e(\mathbf{x})$, and denote this modified vector by $\tilde{\mathbf{y}}_{\mathbf{x}}$. Let $\tilde{K} = \{(\mathbf{x}, \tilde{\mathbf{y}}_{\mathbf{x}}); \mathbf{x} \in \mathcal{X}, \mathbf{y} \in K_{\mathbf{x}}\}$.

Note that the process $(\mathbf{X}(t), \tilde{\mathbf{Y}}(t))$ is a continuous time Markov process with state space $\tilde{K}$. Its macro state transition kernel $\tilde{Q}$ is unchanged for this process. Let $\tilde{\mathcal{A}}_+$ be the restriction of the weak generator $\mathcal{A}_+$on $\mathcal{M}_b^1(\tilde{K})$ for the PDMP $(\mathbf{X}(t), \mathbf{Y}(t))$. That is, for $\tilde{f} \in \mathcal{M}_b^1(\tilde{K})$ and $\tilde{f}_K(\mathbf{x},\mathbf{y}) \equiv \tilde{f}(\mathbf{x}, \tilde{\mathbf{y}}_{\mathbf{x}})$,

$$\tilde{\mathcal{A}}_+ \tilde{f}(\mathbf{x}, \tilde{\mathbf{y}}_{\mathbf{x}}) = \mathcal{A}_+ \tilde{f}_K(\mathbf{x}, \mathbf{y}), \qquad (\mathbf{x}, \mathbf{y}) \in K. \tag{4.51}$$

The following fact is intuitively clear, but its proof clarifies the role of the stationary equation (4.47) of Theorem 4.8.

Theorem 4.9. Let the PDMP $(\mathbf{X}(t), \mathbf{Y}(t))$ have weak kernel $\mathcal{A}_+$ and jump transition kernel Q. Assume that this PDMP satisfies the assumption (4.13a) and the mean lifetime of the i-th continuous component is $1/\mu_i(\mathbf{x})$ for $i \in J_e(\mathbf{x})$. Then, $\tilde{\nu}$ is the stationary distribution of $(\mathbf{X}(t), \tilde{\mathbf{Y}}(t))$ that has a finite intensity for the embedded point process generated by macro state transitions if and only if there exists a finite measure $\tilde{\nu}_{\mathbf{x}}$ on $(\tilde{K}_{\mathbf{x}}, \mathcal{B}(\tilde{K}_{\mathbf{x}}))$ for each $\mathbf{x} \in \mathcal{X}$ such that

$$\lambda \tilde{\nu}_0(\{\mathbf{x}\} \times d\tilde{\mathbf{y}}_{\mathbf{x}}) = \tilde{\nu}_{\mathbf{x}}(d\tilde{\mathbf{y}}_{\mathbf{x}}) + \sum_{i \in J_e(\mathbf{x})} c_{\mathbf{x}i} \mu_i(\mathbf{x}) \tilde{\nu}(\{\mathbf{x}\} \times d\tilde{\mathbf{y}}_{\mathbf{x}}), \qquad \mathbf{x} \in \mathcal{X}, \tag{4.52}$$

$$\int_{\tilde{K}} \tilde{\mathcal{A}}_+ \tilde{f}(\tilde{\mathbf{z}}) \tilde{\nu}(d\tilde{\mathbf{z}}) = \lambda \int_{\partial \tilde{K}} \left(\tilde{f}(\tilde{\mathbf{z}}) - Q\tilde{f}(\tilde{\mathbf{z}}) \right) \tilde{\nu}_0(d\tilde{\mathbf{z}}), \qquad \tilde{f} \in \mathcal{M}_b^1(\tilde{K}), \tag{4.53}$$

where $\tilde{\mathcal{A}}_+$ is the weak generator of $(\mathbf{X}(t), \tilde{\mathbf{Y}}(t))$, that is given by (4.51).

Proof. For necessity, (4.52) is immediate from the decomposition formula of Palm distributions while (4.53) is obtained from Theorem 4.8 and (4.51). To prove sufficiency, we let $\tilde{f}(\tilde{\mathbf{x}}, \mathbf{y}) = 1(\tilde{\mathbf{x}} = \mathbf{x})\tilde{g}(\tilde{\mathbf{y}}_{\mathbf{x}})$ in (4.53). Then, with help of (4.52), we have

$$\begin{aligned} \int_{\tilde{K}_{\mathbf{x}}} \tilde{\mathcal{A}}_+ \tilde{g}(\mathbf{x}, \tilde{\mathbf{y}}_{\mathbf{x}}) \tilde{\nu}(\{\mathbf{x}\} \times d\tilde{\mathbf{y}}_{\mathbf{x}}) = & \int_{\partial \tilde{K}_{\mathbf{x}}} \tilde{g}(\tilde{\mathbf{y}}_{\mathbf{x}}) \tilde{\nu}_{\mathbf{x}}(d\tilde{\mathbf{y}}_{\mathbf{x}}) \\ & + \sum_{i \in J_e(\mathbf{x})} c_{\mathbf{x}i} \mu_i(\mathbf{x}) \int_{\partial \tilde{K}_{\mathbf{x}}} \tilde{g}(\tilde{\mathbf{y}}_{\mathbf{x}}) \tilde{\nu}(\{\mathbf{x}\} \times d\tilde{\mathbf{y}}_{\mathbf{x}}) \\ & - \lambda \int_{\partial \tilde{K}} Q(\tilde{\mathbf{z}}', \{\mathbf{x}\} \times d\tilde{\mathbf{y}}_{\mathbf{x}}) \tilde{g}(\tilde{\mathbf{y}}_{\mathbf{x}}) \tilde{\nu}_0(d\tilde{\mathbf{z}}'). \end{aligned} \tag{4.54}$$

Multiply both sides of this equation by $\prod_{j \in J_e(\mathbf{x})} \frac{\mu_j(\mathbf{x})}{\mu_j(\mathbf{x}) + \theta_j}$, which is the joint Laplace transform of independent and exponentially distributed random variables with means $1/\mu_j(\mathbf{x})$, where θ_j is a nonnegative number, and should not be confused with the shift operator θ_t. Then, the second term in the right-hand side can be computed as

$$\begin{aligned}
&\sum_{i \in J_e(\mathbf{x})} c_{\mathbf{x}i}\mu_i(\mathbf{x}) \int_{\partial \tilde{K}_\mathbf{x}} \tilde{g}(\tilde{\mathbf{y}}_\mathbf{x})\tilde{\nu}(\{\mathbf{x}\} \times d\tilde{\mathbf{y}}_\mathbf{x}) \prod_{j \in J_e(\mathbf{x})} \frac{\mu_j(\mathbf{x})}{\mu_j(\mathbf{x}) + \theta_j} \\
&= \sum_{i \in J_e(\mathbf{x})} c_{\mathbf{x}i}\mu_i(\mathbf{x}) \left(1 - \frac{\theta_i}{\mu_i(\mathbf{x}) + \theta_i}\right) \int_{\partial \tilde{K}_\mathbf{x}} \tilde{g}(\tilde{\mathbf{y}}_\mathbf{x})\tilde{\nu}(\{\mathbf{x}\} \times d\tilde{\mathbf{y}}_\mathbf{x}) \prod_{j \in J_e(\mathbf{x})\setminus\{i\}} \frac{\mu_j(\mathbf{x})}{\mu_j(\mathbf{x}) + \theta_j} \\
&= \sum_{i \in J_e(\mathbf{x})} c_{\mathbf{x}i}\mu_i(\mathbf{x}) \int_{\partial \tilde{K}_\mathbf{x}} \tilde{g}(\tilde{\mathbf{y}}_\mathbf{x})\tilde{\nu}(\{\mathbf{x}\} \times d\tilde{\mathbf{y}}_\mathbf{x}) \prod_{j \in J_e(\mathbf{x})\setminus\{i\}} \frac{\mu_j(\mathbf{x})}{\mu_j(\mathbf{x}) + \theta_j} \\
&\qquad - \sum_{i \in J_e(\mathbf{x})} c_{\mathbf{x}i}\theta_i \int_{\partial \tilde{K}_\mathbf{x}} \tilde{g}(\tilde{\mathbf{y}}_\mathbf{x})\tilde{\nu}(\{\mathbf{x}\} \times d\tilde{\mathbf{y}}_\mathbf{x}) \prod_{j \in J_e(\mathbf{x})} \frac{\mu_j(\mathbf{x})}{\mu_j(\mathbf{x}) + \theta_j}.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&\int_{\tilde{K}_\mathbf{x}} \tilde{\mathcal{A}}_+\tilde{g}(\mathbf{y})\tilde{\nu}(\{\mathbf{x}\} \times d\tilde{\mathbf{y}}_\mathbf{x}) \prod_{j \in J_e(\mathbf{x})} \frac{\mu_j(\mathbf{x})}{\mu_j(\mathbf{x}) + \theta_j} \\
&\qquad + \sum_{i \in J_e(\mathbf{x})} c_{\mathbf{x}i}\theta_i \int_{\partial \tilde{K}_\mathbf{x}} \tilde{g}(\tilde{\mathbf{y}}_\mathbf{x})\tilde{\nu}(\{\mathbf{x}\} \times d\tilde{\mathbf{y}}_\mathbf{x}) \prod_{j \in J_e(\mathbf{x})} \frac{\mu_j(\mathbf{x})}{\mu_j(\mathbf{x}) + \theta_j} \\
&= \int_{\partial \tilde{K}_\mathbf{x}} \tilde{g}(\tilde{\mathbf{y}}_\mathbf{x})\tilde{\nu}_\mathbf{x}(d\tilde{\mathbf{y}}_\mathbf{x}) \prod_{j \in J_e(\mathbf{x})} \frac{\mu_j(\mathbf{x})}{\mu_j(\mathbf{x}) + \theta_j} \\
&\qquad + \sum_{i \in J_e(\mathbf{x})} c_{\mathbf{x}i}\mu_i(\mathbf{x}) \int_{\partial \tilde{K}_\mathbf{x}} \tilde{g}(\tilde{\mathbf{y}}_\mathbf{x})\tilde{\nu}(\{\mathbf{x}\} \times d\tilde{\mathbf{y}}_\mathbf{x}) \prod_{j \in J_e(\mathbf{x})\setminus\{i\}} \frac{\mu_j(\mathbf{x})}{\mu_j(\mathbf{x}) + \theta_j} \\
&\qquad - \lambda \int_{\partial \tilde{K}} Q(\tilde{\mathbf{z}}', \{\mathbf{x}\} \times d\tilde{\mathbf{y}}_\mathbf{x})\tilde{g}(\tilde{\mathbf{y}}_\mathbf{x})\tilde{\nu}_0(d\tilde{\mathbf{z}}') \prod_{j \in J_e(\mathbf{x})} \frac{\mu_j(\mathbf{x})}{\mu_j(\mathbf{x}) + \theta_j}.
\end{aligned}$$

This is identical with (4.47) with f given by

$$f(\mathbf{x}', \mathbf{y}) = 1(\mathbf{x}' = \mathbf{x})\tilde{g}(\tilde{\mathbf{y}}_\mathbf{x}) \prod_{j \in J_e(\mathbf{x})} e^{-\theta_i y_i}.$$

Since θ_i can be any positive number, this class of function f is sufficiently large to determine a distribution on K. Hence,

$$\nu(\{\mathbf{x}\} \times d\mathbf{y}) = \tilde{\nu}(\{\mathbf{x}\} \times d\tilde{\mathbf{y}}_\mathbf{x}) \prod_{j \in J_e(\mathbf{x})} \mu_j(\mathbf{x})e^{-\mu_j(\mathbf{x})y_j} dy_j$$

is the stationary distribution of the PDMP $(\mathbf{X}(t), \mathbf{Y}(t))$, so $\tilde{\nu}$ is that of $(\mathbf{X}(t), \tilde{\mathbf{Y}}(t))$. □

It is notable that (4.52) is necessary to get the stationary distribution. Of course, we can combine (4.52) and (4.53) substituting the former into the latter.

Example 4.14 (State dependent workload, revisited). In Example 4.13, we formulate the workload process $V(t)$ of Example 4.8 for the $M/G/1$ queue with state dependent processing rate r as the PDMP $(X(t), (Y_1(t), Y_2(t)))$. Since $T_n - T_{n-1}$ is exponentially distributed and independent of everything else, we can drop $Y_1(t)$, and Theorem 4.9 is applicable. Although this is not so much helpful to find the stationary distribution since (4.52) and (4.53) are equivalent to (4.32), we can see how Theorem 4.9 is applied. □

4.14 GSMP and RGSMP

In many queueing applications of the PDMP, all the continuous components $Y_\ell(t)$ count the remaining lifetimes, and the macro state transitions due to Q is independent of non-zero remaining lifetimes. Assume that the remaining lifetimes decrease with constant rates. In this case, the weak generator $\mathcal{A}_+$ has a simpler form:

$$\mathcal{A}_+ f(\mathbf{x},\mathbf{y}) = -\sum_{i=1}^{m(\mathbf{x})} c_{\mathbf{x}i} \frac{\partial}{\partial y_i} f(\mathbf{x},\mathbf{y}), \tag{4.55}$$

where $c_{\mathbf{x}i}$ are nonnegative constants for each $\mathbf{x}$ and i. Furthermore, Q is also simpler. We introduce this class of models adding more structure to the macro states.

Definition 4.11 (GSMP). Let $\mathcal{X}$ and $\mathcal{S}$ be countable or finite sets. Their elements are called a macro state and a site, respectively. For each $\mathbf{x} \in \mathcal{X}$, a finite and non-empty subset of $\mathcal{S}$ is associated, and denoted by $A(\mathbf{x})$, whose element is called an active site under macro state $\mathbf{x}$. For each $s \in A(\mathbf{x})$, a clock is attached, and counts its remaining life time r_s. Let $r(\mathbf{x}) = \{(s,r_s); s \in A(\mathbf{x})\}$.

Assume the following dynamics of macro states and clocks.

(4.14a) Under macro state $\mathbf{x}$, the clock at site $s \in A(\mathbf{x})$ advances with speed $\overline{c}_{\mathbf{x}s}$.
(4.14b) If the remaining lifetime of clocks at sites in $U \subset A(\mathbf{x})$ simultaneously expire under macro state $\mathbf{x}$, then the macro state changes to $\mathbf{x}'$ with probability $p_U(\mathbf{x},\mathbf{x}')$.
(4.14c) Under the above transition, the remaining lifetimes of clocks at sites $A(\mathbf{x}) \setminus U$ are retained, and new clocks are activates on sites $A(\mathbf{x}') \setminus A(\mathbf{x})$ with lifetimes independently sampled from the distribution determined by their sites and new macro state $\mathbf{x}'$.

Thus, $A(\mathbf{x}) \setminus U$ must be a subset of $A(\mathbf{x}')$. Let $\mathbf{X}(t)$ be a macro state at time t, and let $R_s(t)$ be remaining lifetime of the clock at site $s \in A(\mathbf{x})$ at time t. Then, $(\mathbf{X}(t), \{R_s(t); s \in A(\mathbf{x})\})$ is a Markov process. We refer to this Markov process as a generalized semi-Markov process, GSMP for short, with macro state space $\mathcal{X}$ and site space $\mathcal{S}$. □

This GSMP is not exactly the PDMP (piece-wise deterministic Markov process), but can be reduced to it. To see this, let $m(\mathbf{x})$ be the number of elements of $A(\mathbf{x})$, and let $J(\mathbf{x}) = \{1,2,\ldots,m(\mathbf{x})\}$, where $m(\mathbf{x})$ is a finite positive integer by the assumption on $A(\mathbf{x})$. For each $\mathbf{x} \in \mathcal{X}$, define one to one mapping $\xi_{\mathbf{x}}$ from $A(\mathbf{x})$ to $J(\mathbf{x})$. For each $\ell \in J(\mathbf{x})$, let $y_\ell = v_{\xi_{\mathbf{x}}^{-1}(\ell)}$. Thus, site $s \in A(\mathbf{x})$ is mapped to $\ell \in J(\mathbf{x})$ with the remaining lifetime of the clock attached to s. Let $K_{\mathbf{x}} = [0,\infty)^{m(\mathbf{x})}$ and $K = \cup_{\mathbf{x}\in\mathcal{X}}\{\mathbf{x}\} \times K_{\mathbf{x}}$. Let $R_s(t)$ be the remaining lifetime of the clock at site s, and with $Y_\ell(t) = R_{\xi^{-1}_{\mathbf{X}(t)}(\ell)}(t)$ let

$$\mathbf{Y}(t) = (Y_1(t), Y_2(t), \ldots, Y_{m(\mathbf{X}(t))}(t)).$$

We also define the jump transition kernel Q as, for $\mathbf{x},\mathbf{x}' \in \mathcal{X}$, $\mathbf{y} \in K_{\mathbf{x}}$ and $B_\ell \in \mathcal{B}(\mathbb{R})$,

$$\begin{aligned} &Q((\mathbf{x},\mathbf{y}), \{\mathbf{x}'\} \times B_1 \times \cdots \times B_{m(\mathbf{x}')}) \\ &\qquad = p_U(\mathbf{x},\mathbf{x}') \prod_{\ell\in\xi(A(\mathbf{x})\setminus U)} 1(y_\ell \in B_\ell) \prod_{\ell\in\xi(A(\mathbf{x}')\setminus A(\mathbf{x}))} F_{\mathbf{x}'\xi^{-1}_{\mathbf{x}'}(\ell)}(B_\ell), \end{aligned}$$

where U is the set of all expiring sites under $\mathbf{x}$ and the remaining lifetimes $y_{\xi^{-1}_{\mathbf{x}}(s)}$ of the clock at site $s \in A(\mathbf{x})$, and $F_{\mathbf{x}s}$ is the new lifetime distribution of the clock at site s under macro state $\mathbf{x}$. Note that $F(B)$ is defined for distribution F as

$$F(B) = \int_B F(du), \qquad B \in \mathcal{B}(\mathbb{R}).$$

Then, we have PDMP $\{(\mathbf{X}(t),\mathbf{Y}(t))\}$ with state space K and jump transition kernel Q. We refer to $\{(\mathbf{X}(t),\mathbf{Y}(t))\}$ as a canonical form of GSMP.

From the assumption on the speed of clocks, we have

$$\frac{dY_\ell(t)}{dt} = -c_{\mathbf{X}(t)\ell}, \qquad \ell \in J(\mathbf{X}(t)),$$

where $c_{\mathbf{x}\ell} = \overline{c}_{\mathbf{x}\xi^{-1}(\ell)}$. Hence, let $\mathbf{y} = (y_1,\ldots,y_{m(\mathbf{x})})$, then

$$\zeta(\mathbf{x},\mathbf{y}) = \min\left\{\frac{y_1}{c_{\mathbf{x}1}},\ldots,\frac{y_{m(\mathbf{x})}}{c_{\mathbf{x}m(\mathbf{x})}}\right\},$$

$$\psi(\mathbf{x},\mathbf{y}) = \left(\frac{y_1}{c_{\mathbf{x}1}} - \zeta(\mathbf{x},\mathbf{y}),\ldots,\frac{y_{m(\mathbf{x})}}{c_{\mathbf{x}m(\mathbf{x})}} - \zeta(\mathbf{x},\mathbf{y})\right).$$

Many queueing models and their networks can be described by GSMP. For those models, sites correspond with arrivals and services, and the remaining lifetimes are the remaining arrival times and the remaining workloads. Particularly, GSMP is useful for those queues with the first-come and first-served discipline since sites for service are unchanged for them.

However, GSMP is not so convenient when services are interrupted. In this case, we have to keep track of the remaining workloads of all customers who once started service. Then, the macro state has to accommodate all the sites as they are since clocks are fixed at sites in GSMP. This often unnecessarily complicates analysis, particularly when the number of active sites is unbounded.

To reduce this complication, one may think of reallocating clocks on sites at each transition instants. This is basically equivalent to only work on the canonical form with the reallocation. Let us define this model.

Definition 4.12 (RGSMP). Let $\mathcal{X}$ be a finite or countable set for a macro state space, and let $J(\mathbf{x}) \equiv \{1,2,\ldots,m(\mathbf{x})\}$ of the set of all active sites under macro state $\mathbf{x}$. Let D be the index set of lifetime distributions for clocks at their activation, where the same distribution may have different indexes. An active clock is allocated to each element of $J(\mathbf{x})$ and has the remaining lifetime, but this allocation may change at the macro state transitions in the following way.

(4.14d) Under macro state $\mathbf{x}$, the remaining lifetime of the clock at site $\ell \in J(\mathbf{x})$ decreases with rate $c_{\mathbf{x}\ell}$, where there is at least one positive rate.

(4.14e) Each clock at site $\ell \in J(\mathbf{x})$ has an index in D. Denote this index by $\gamma_{\mathbf{x}}(\ell)$. This $\gamma_{\mathbf{x}}$ is a mapping from $J(\mathbf{x})$ to D, which is not necessarily one-to-one.

(4.14f) When all clocks of sites in set U simultaneously expire, macro state $\mathbf{x}$ changes to $\mathbf{x}'$ activating clocks on sites in set U' with probability $p((\mathbf{x},U),(\mathbf{x}',U'))$.

(4.14g) At this macro state transition, clocks on $J(\mathbf{x}) \setminus U$ are reallocated on $J(\mathbf{x}') \setminus U'$ by one-to-one mapping $\Gamma_{\mathbf{x}U,\mathbf{x}'U'}$ onto $J(\mathbf{x}') \setminus U'$, whose domain $\Gamma^{-1}_{\mathbf{x}U,\mathbf{x}'U'}(J(\mathbf{x}') \setminus U')$ is a subset of $J(\mathbf{x}) \setminus U$. The clocks at sites in the set:

$$(J(\mathbf{x}) \setminus U) \setminus \Gamma^{-1}_{\mathbf{x}U,\mathbf{x}'U'}(J(\mathbf{x}') \setminus U')$$

are said to be interrupted. Under this reallocation, the remaining lifetimes of the reallocated clocks and their indexes are unchanged while newly activated clocks with indexes $d \in D$ have the lifetimes independently sampled subject to distribution F_d's.

Let $\mathbf{X}(t)$ and $\mathbf{Y}(t)$ be the macro state and the remaining lifetime vector at time t. Then, $\{(\mathbf{X}(t),\mathbf{Y}(t))\}$ is the PDMC, and we refer to it as a reallocatable generalized semi-Markov process, RGSMP for short. □

Note that the transition kernel Q at macro state transitions is given by

$$\begin{aligned} &Q((\mathbf{x},\mathbf{y}),(\mathbf{x}',B_1 \times \cdots \times B_{m(\mathbf{x}')})) \\ &\quad = p((\mathbf{x},U),(\mathbf{x}',U')) \prod_{j \in \Gamma_{\mathbf{x}U\mathbf{x}'}(J(\mathbf{x})\setminus U)} 1(y_j \in B_j) \prod_{j \in U'} F_{\gamma_{\mathbf{x}'}(j)}(B_j). \end{aligned}$$

We assume that $\Gamma_{\mathbf{x}U,\mathbf{x}'U'}$ is deterministic for simplicity, but it could be random without any difficulty. In applications, U is usually a singleton, that is, $U = \{\ell\}$ for some ℓ. In this case, $p((\mathbf{x},U),(\mathbf{x}',U'))$ and $\Gamma_{\mathbf{x}U,\mathbf{x}'U'}$ are simply written as $p((\mathbf{x},\ell),(\mathbf{x}',U'))$ and $\Gamma_{\mathbf{x}\ell,\mathbf{x}'U'}$, respectively.

Although the canonical form of the remaining lifetimes is sufficient for RGSMP, it may not be always convenient. For example, if there are different groups of sites and reallocations are only taken within each group, then a finite set of multidimensional vectors is more convenient than one multidimensional vector $\mathbf{y}$. In this case, $J(\mathbf{x})$ is divided into subsets $J_{s_1}(\mathbf{x}),\ldots,J_{s_k}(\mathbf{x})$ for the number k of such groups and their indexes $s_1,\ldots,s_k$. Similarly, $\gamma_{\mathbf{x}}(\ell)$ is divided into $\gamma_{s_1\mathbf{x}}(\ell_1),\ldots,\gamma_{s_k\mathbf{x}}(\ell_k)$.

In our definition of RGSMP, some of active clocks may expire at the transition. In queueing applications, this may be the case that a customer being serviced is forced to leave. For example, so called a negative customer causes such an event.

We now specialize 4.6 to RGSMP (reallocatable generalized semi-Markov process) of Definition 4.12. In this case, Laplace transform is convenient.

Corollary 4.7. Assume RGSMP satisfies the assumptions in Theorem 4.8. For each $\mathbf{x} \in \mathcal{X}$, let $\theta_{\mathbf{x}} = (\theta_1, \ldots, \theta_{m(\mathbf{x})})$, $\theta_\ell \geq 0$, $\langle \theta_{\mathbf{x}}, \mathbf{y} \rangle = \sum_{\ell=1}^{m(\mathbf{x})} \theta_\ell y_\ell$, then (4.50) can be replaced by

$$\sum_{\ell=1}^{m(\mathbf{x})} c_{\mathbf{x}\ell} \theta_\ell \hat{\nu}(\mathbf{x}, \theta_{\mathbf{x}}) = \lambda \left(\hat{\nu}_0(\mathbf{x}, \theta_{\mathbf{x}}) - \hat{\nu}_0^+(\mathbf{x}, \theta_{\mathbf{x}}) \right), \qquad \theta \geq 0, \mathbf{x} \in \mathcal{X}, \qquad (4.56)$$

where

$$\hat{\nu}(\mathbf{x}, \theta_{\mathbf{x}}) = \int_{K_{\mathbf{x}+}} e^{-\langle \theta_{\mathbf{x}}, \mathbf{y} \rangle} \nu(\mathbf{x}, d\mathbf{y}),$$
$$\hat{\nu}_0(\mathbf{x}, \theta_{\mathbf{x}}) = \int_{K_{\mathbf{x}0}} e^{-\langle \theta_{\mathbf{x}}, \mathbf{y} \rangle} \nu_0(\mathbf{x}, d\mathbf{y}),$$
$$\hat{\nu}_0^+(\mathbf{x}, \theta_{\mathbf{x}}) = \int_{K_0} \int_{K_{\mathbf{x}+}} e^{-\langle \theta_{\mathbf{x}}, \mathbf{y} \rangle} Q(\mathbf{z}', (\mathbf{x}, d\mathbf{y})) \nu_0(\mathbf{z}').$$

Since function $e^{-\langle \theta_{\mathbf{x}}, \mathbf{y} \rangle}$ is in $C_b^1(K)$, the necessity is immediate. For the necessity, we need to approximate functions in C_b^1 with compact supports by Fourier series. This can be found in [26] again.

Up to now, we are mainly concerned with a single point process N for the macro state transitions. We can decompose this N into point processes observed at sites. For each subset U of $\{1, 2, \ldots\}$, define point process N_U as

$$N_U(B) = \sum_{t \in B} \sum_{\mathbf{x} \in \mathcal{X}} 1(\mathbf{X}(t) = \mathbf{x}, Y_\ell(t) = 0, \ell \in U \cap J(\mathbf{x})), \qquad B \in \mathcal{B}(\mathbb{R}),$$

and let $\lambda_U = E(N_U((0, 1]))$. Since $\lambda_U \leq \lambda < \infty$, we can define Palm distribution P_U of P with respect to N_U. Denote the distribution of $\mathbf{Z}(t)$ under P_U by ν_U. Let $\hat{\nu}_U(\mathbf{x}, \theta_{\mathbf{x}})$ be the Laplace transform with respect to the remaining lifetimes under macro state $\mathbf{x} \in \mathcal{X}$, and let

$$\hat{\nu}_U^+(\mathbf{x}, \theta_{\mathbf{x}}) = \int_{K_0} \int_{K_{\mathbf{x}+}} e^{-\langle \theta_{\mathbf{x}}, \mathbf{y} \rangle} Q(\mathbf{z}', (\mathbf{x}, d\mathbf{y})) \nu_U(\mathbf{z}').$$

Then, (4.56) can be replaced by

$$\sum_{\ell=1}^{m(\mathbf{x})} c_{\mathbf{x}\ell} \theta_\ell \hat{\nu}(\mathbf{x}, \theta_{\mathbf{x}}) = \sum_U \lambda_U \left(\hat{\nu}_U(\mathbf{x}, \theta_{\mathbf{x}}) - \hat{\nu}_U^+(\mathbf{x}, \theta_{\mathbf{x}}) \right), \qquad \theta \geq 0. \qquad (4.57)$$

In many cases, U is a singleton for $\lambda_U > 0$. In this case, we simply write $\lambda_{\{\ell\}}$ and $\nu_{\{\ell\}}$ as λ_ℓ and ν_ℓ, respectively, for $U = \{\ell\}$.

4.15 Exponential and non-exponential clocks in RGSMP

In this section, we consider the stationary distribution of RGSMP (reallocatable generalize semi-Markov process), provided it exists. In what follows, we use the notations in Definition 4.12, and assume that $\{(\mathbf{X}(t),\mathbf{Y}(t))\}$ is stationary under P.

As we have considered in Section 4.13, it is interesting to see the case where some of lifetime distributions are exponential. We here consider such a case for RGSMP. Denote the set of the indexes in D which specify the exponential distributions by D_e. Similarly, let $J_{\mathrm{e}}(\mathbf{x})$ be the set of the sites whose clocks have indexes in D_e. Those clocks are activated with lifetimes subject to the exponential distributions. For the other distributions, we let

$$D_{\mathrm{g}} = D \setminus D_{\mathrm{e}}, \qquad J_{\mathrm{g}}(\mathbf{x}) = J(\mathbf{x}) \setminus J_{\mathrm{e}}(\mathbf{x}).$$

In this section, we shall use Theorem 4.9 and 4.7 to characterize the stationary distribution. We first prepare some notations. For each $d \in D$, denote the mean of distribution F_d by m_d, and its reciprocal by μ_d. Denote the Laplace transform of F_d by $\hat{F}_d(\theta)$. Since F_d is exponential for $d \in D_{\mathrm{e}}$,

$$F_d(x) = 1 - e^{-\mu_d x}, \quad x \geq 0, \qquad\qquad \hat{F}_d(\theta) = \frac{\mu_d}{\mu_d + \theta}, \quad \theta \geq 0.$$

Let $\theta_{\mathbf{x}} = (\theta_1, \ldots, \theta_{m(\mathbf{x})})$. For $U \subset J(\mathbf{x})$, let $\theta_{\mathbf{x}}(U)$ denote the $\theta_{\mathbf{x}}$ in which the components with indexes in U is replaced by 0. In particular, if $U = \{\ell\}$, then $\theta_{\mathbf{x}}(U)$ is denoted by $\theta_{\mathbf{x}}(\ell)$. Let N_ℓ be the point process generated by expiring instants of clocks at site ℓ. This point process is obviously stationary under P. In what follows, we also assume

(4.15a) The mean m_d of F_d is finite for all $d \in D$.
(4.15b) Not more than one clock simultaneously expires.
(4.15c) $\sum_{\ell=1}^{\infty} \lambda_\ell < \infty$, where λ_ℓ is the intensity of N_ℓ.

Let P_ℓ be the Palm distribution concerning N_ℓ. We denote the distribution of $(\mathbf{X}(t),\mathbf{Y}(t))$ under P by ν, and its Laplace transform concerning $\mathbf{Y}(t)$ under $\mathbf{X}(t) = \mathbf{x}$ by $\hat{\nu}(\mathbf{x},\theta_{\mathbf{x}})$ for each $\mathbf{x} \in \mathcal{X}$. Similarly, the distribution of $(\mathbf{X}(0-),\mathbf{Y}(0-))$ under the Palm distribution P_ℓ and its Laplace transform under $\mathbf{X}(0-) = \mathbf{x}$ are denoted by ν_ℓ and $\hat{\nu}_\ell(\mathbf{x},\theta_{\mathbf{x}})$, respectively.

Lemma 4.12. For $\mathbf{x} \in \mathcal{X}$ and $\theta_{\mathbf{x}} \geq 0$, we have

$$\hat{\nu}(\mathbf{x},\theta_{\mathbf{x}}) = \hat{\nu}(\mathbf{x},\theta_{\mathbf{x}}(J_{\mathrm{e}}(\mathbf{x}))) \prod_{i \in J_{\mathrm{e}}(\mathbf{x})} \frac{\mu_{\gamma_{\mathbf{x}}(i)}}{\mu_{\gamma_{\mathbf{x}}(i)} + \theta_i}, \tag{4.58}$$

$$\hat{\nu}_\ell(\mathbf{x},\theta_{\mathbf{x}}) = \hat{\nu}_\ell(\mathbf{x},\theta_{\mathbf{x}}(J_{\mathrm{e}}(\mathbf{x}))) \prod_{i \in J_{\mathrm{e}}(\mathbf{x}) \setminus \{\ell\}} \frac{\mu_{\gamma_{\mathbf{x}}(i)}}{\mu_{\gamma_{\mathbf{x}}(i)} + \theta_i}, \tag{4.59}$$

$$c_{\mathbf{x}\ell}\mu_{\gamma_{\mathbf{x}}(\ell)} \hat{\nu}(\mathbf{x},\theta_{\mathbf{x}}(J_{\mathrm{e}}(\mathbf{x}))) = \lambda_\ell \hat{\nu}_\ell(\mathbf{x},\theta_{\mathbf{x}}(J_{\mathrm{e}}(\mathbf{x}))), \qquad \ell \in J_{\mathrm{e}}(\mathbf{x}). \tag{4.60}$$

Proof. (4.58) and (4.59) are immediate from the memoryless property of the exponential distribution. Substituting them into (4.57) and letting $\theta_\ell \to \infty$ yield (4.60). □

The next result is a specialization of 4.7 to the case that some of lifetime distributions are exponential, but can be viewed as a special case of Theorem 4.9.

Theorem 4.10. Under the assumptions (4.15a), (4.15b) and (4.15c), RGSMP has the stationary distribution if and only if there exist Laplace transforms $\hat{\nu}, \hat{\nu}_\ell$ and λ_ℓ ($\ell = 1, 2, \ldots$) such that (4.60) holds and, for each $\mathbf{x} \in \mathcal{X}$ and $\theta_{\mathbf{x}}(J_{\mathrm{e}}(\mathbf{x})) \geq 0$,

$$\begin{aligned}
&\sum_{i \in J_{\mathrm{g}}(\mathbf{x})} c_{\mathbf{x}i} \theta_i \hat{\nu}(\mathbf{x}, \theta_{\mathbf{x}}(J_{\mathrm{e}}(\mathbf{x}))) \\
&\quad = \sum_{i \in J(\mathbf{x})} \lambda_i \hat{\nu}_i(\mathbf{x}, \theta_{\mathbf{x}}(J_{\mathrm{e}}(\mathbf{x}))) - \sum_{\mathbf{x}' \in \mathcal{X}} \sum_{i \in J(\mathbf{x}')} \sum_{U \subset J(\mathbf{x})} \lambda_i \hat{\nu}_i(\mathbf{x}', \hat{\Gamma}^{-1}_{\mathbf{x}'i,\mathbf{x}U}(\theta_{\mathbf{x}}(J_{\mathrm{e}}(\mathbf{x}))) \\
&\qquad\qquad \times p((\mathbf{x}', i), (\mathbf{x}, U)) \prod_{j \in U \cap J_{\mathrm{g}}(\mathbf{x})} \hat{F}_{\gamma_{\mathbf{x}}(j)}(\theta_j),
\end{aligned} \tag{4.61}$$

where $\hat{\Gamma}^{-1}_{\mathbf{x}'i,\mathbf{x}U}(\theta_{\mathbf{x}})$ is the $m(\mathbf{x}')$-dimensional vector whose j-th entry is $\theta_{\Gamma_{\mathbf{x}'i,\mathbf{x}U}(j)}$ if $j \neq i$ and $\Gamma_{\mathbf{x}'i,\mathbf{x}U}(j) \in J(\mathbf{x})$ and equals 0 otherwise. In this case, $\hat{\nu}$ is the Laplace transform of the stationary distribution ν.

Remark 4.7. It is not hard to see that (4.61) is a special case of (4.54).

Proof. We apply 4.7. From the assumption (4.15a),

$$\lambda \hat{\nu}_0(\mathbf{x}, \theta_{\mathbf{x}}) = \sum_{\ell=1}^{\infty} \lambda_\ell \hat{\nu}_\ell(\mathbf{x}, \theta_{\mathbf{x}}).$$

Similarly, from the definition of $\hat{\Gamma}^{-1}_{\mathbf{x}'i\mathbf{x}U}$,

$$\begin{aligned}
\lambda \hat{\nu}_0^+(\mathbf{x}, \theta_{\mathbf{x}}) &= \sum_{i=1}^{\infty} \lambda_i \hat{\nu}_i^+(\mathbf{x}, \theta_{\mathbf{x}}) \\
&= \sum_{i=1}^{\infty} \sum_{\mathbf{x}' \in \mathcal{X}} \sum_{U \subset J(\mathbf{x})} \lambda_i \hat{\nu}_i^-(\mathbf{x}', \hat{\Gamma}^{-1}_{\mathbf{x}'i,\mathbf{x}U}(\theta_{\mathbf{x}'})) p((\mathbf{x}', i), (\mathbf{x}, U)) \prod_{j \in U \cap J_{\mathrm{g}}(\mathbf{x})} \hat{F}_{\gamma_{\mathbf{x}}(j)}(\theta_j).
\end{aligned}$$

Substituting these formulas together with (4.58) and (4.59) into (4.56) and dividing both sides by $\prod_{j \in J_{\mathrm{e}}(\mathbf{x})} \frac{\mu_{\gamma_{\mathbf{x}}(j)}}{\mu_{\gamma_{\mathbf{x}}(j)} + \theta_j}$, we have

$$
\begin{aligned}
&\sum_{i=1}^{m(\mathbf{x})} c_{\mathbf{x}i}\theta_i \hat{v}(\mathbf{x}, \theta_{\mathbf{x}}(J_{\mathrm{e}}(\mathbf{x}))) \\
&\quad = \sum_{i\in J_{\mathrm{g}}(\mathbf{x})} \lambda_i \hat{v}_i(\mathbf{x}, \theta_{\mathbf{x}}(J_{\mathrm{e}}(\mathbf{x})) + \sum_{i\in J_{\mathrm{e}}(\mathbf{x})} \frac{\lambda_i(\mu_{\gamma_{\mathbf{x}}(i)} + \theta_i)}{\mu_{\gamma_{\mathbf{x}}(i)}} \hat{v}_i(\mathbf{x}, \theta_{\mathbf{x}}(J_{\mathrm{e}}(\mathbf{x})) \\
&\qquad - \sum_{\mathbf{x}'\in\mathcal{X}} \sum_{i=1}^{m(\mathbf{x}')} \sum_{U\subset J(\mathbf{x})} \lambda_i \hat{v}_i^-(\mathbf{x}', \hat{\Gamma}_{\mathbf{x}'i,\mathbf{x}U}(\theta_{\mathbf{x}'}(J_{\mathrm{e}}(\mathbf{x}'))))p((\mathbf{x}', i), (\mathbf{x}, U)) \prod_{j\in U\cap J_{\mathrm{g}}(\mathbf{x})} \hat{F}_{\gamma_{\mathbf{x}}(j)}(\theta_j).
\end{aligned}
$$

By 4.12, (4.60) is necessary. We apply it to the second term in the right-hand side of this equation, then we can see that the terms of $i \in J_{\mathrm{e}}(\mathbf{x})$ in the left-hand side are cancelled, which yields (4.61). Thus, (4.60) and (4.61) are necessary. These arguments can be traced back, so the converse is proved (see also the proof of Theorem 4.9). □

Example 4.15. Let us formulate the $M/G/1$ queue of Example 4.7 by the RGSMP. We here assume the first-come first-served discipline. Let $X(t)$ be the number of customers in the system, and $R(t)$ be the remaining service time of a customer in service, where $R(t) = 0$ if the system is empty. Obviously, $(X(t), R(t))$ is a continuous-time Markov chain, and it is easy to see that this process is a RGSMP.

We show how the notations of the RGSMP are specified in this case. Let $\mathcal{X} = \{0, 1, 2, \ldots\}$. $D = \{0, 1\}$, where 0 represents the exponential distribution with mean λ^{-1}, and 1 represents a generic distribution with mean μ^{-1} and distribution F. Define the jump transition function by

$$
\begin{aligned}
&p((n,0),(n+1,U)) = 1 \text{ if } n = 0 \text{ and } U = \{0,1\} \text{ or if } n \geq 1 \text{ and } U = \{0\}, \\
&p((n,1),(n-1,U)) = 1 \text{ if } n = 1 \text{ and } U = \emptyset \text{ or if } n \geq 2 \text{ and } U = \{1\},
\end{aligned}
$$

and let $c_{n,0} = c_{n+1,1} = 1$ for $n \geq 0$. Let

$$
J_e(0) = \{0,1\}, \quad J_e(n) = \{0\}, \quad J_g(0) = \emptyset, \quad J_g(n) = \{1\}, \qquad n \geq 1.
$$

Thus, we indeed have the RGSMP. Assume the stability condition $\rho \equiv \lambda/\mu < 1$. In what follows we solve the stationary equation (4.61), which becomes

$$
\begin{aligned}
&0 = \lambda \hat{v}(0,0) - \lambda \hat{v}(1,0), \\
&\theta \hat{v}(1,\theta) = \lambda(\hat{v}(1,\theta) + \hat{v}(1,0)) - \lambda(\hat{v}(0,0) + \hat{v}(2,0))\hat{F}(\theta), \\
&\theta \hat{v}(n,\theta) = \lambda(\hat{v}(n,\theta) + \hat{v}(n,0)) - \lambda(\hat{v}(n-1,\theta) + \hat{v}(n+1,0)\hat{F}(\theta)),
\end{aligned}
$$

for $n \geq 2$, where we have used the fact that $\lambda_1 = \lambda$. By letting $\theta = 0$ in these formulas. it is easy to see that $\hat{v}_1(n,0) = \hat{v}(n-1,0)$ for $n \geq 1$. Then, it is routine to solve these stationary equations by taking the generating function:

$$
\hat{v}_*(z,\theta) = 1 - \rho + \sum_{n=1}^{\infty} z^n \hat{v}(n,\theta).
$$

Let $\pi(0) = \hat{v}(0,0)$. This yields

$$(\theta-\lambda(1-z))(\hat{v}_*(z,\theta)-\pi(0))=\lambda(1-z)\hat{F}(\theta)\pi(0)+\lambda(z-\hat{F}(\theta))\hat{v}_*(z,0)(4.62)$$

Let $\theta=\lambda(1-z)$ in this equation, then we have

$$\hat{v}_*(z,0)=\frac{(1-\rho)(1-z)\hat{F}(\lambda(1-z))}{\hat{F}(\lambda(1-z))-z}. \tag{4.63}$$

We can compute $\hat{v}_*(z,\theta)$ by substituting this into (4.62). These results are well known. The advantage of the present derivation is that the existence of the density of F is not needed, which is often assumed in the literature. □

If $D_g=\emptyset$ in Theorem 4.10, i.e., all the lifetimes are exponentially distributed, then the set of equations (4.60) and (4.61) with $\theta_i=0$ uniquely determines the stationary distribution of the macro states. Hence, we have the following corollary.

Corollary 4.8. For the RGSMP satisfying (4.15a), (4.15b) and (4.15c), if all the lifetime distributions are exponential, then a probability distribution π on $\mathcal{X}$ is the stationary distribution of $\mathbf{X}(t)$ if and only if, for all $\mathbf{x}\in\mathcal{X}$,

$$\sum_{\ell\in J(\mathbf{x})} c_{\mathbf{x}\ell}\mu_{\gamma_{\mathbf{x}}(\ell)}\pi(\mathbf{x})=\sum_{\mathbf{x}'\in\mathcal{X}}\sum_{\ell\in J(\mathbf{x}')}\sum_{U\subset J(\mathbf{x})} c_{\mathbf{x}'\ell}\mu_{\gamma_{\mathbf{x}'}(\ell)}\pi(\mathbf{x}')p((\mathbf{x}',\ell),(\mathbf{x},U)). \tag{4.64}$$

Equation (4.64) can be interpreted as the stationary equation for the macro state. To see this, define the transition rate function $q(\mathbf{x},\mathbf{x}')$ as

$$q(\mathbf{x},\mathbf{x}')=\sum_{\ell\in J(\mathbf{x}')}\sum_{U\subset J(\mathbf{x})} c_{\mathbf{x}\ell}\mu_{\gamma_{\mathbf{x}}(\ell)}p((\mathbf{x},\ell),(\mathbf{x}',U)).$$

Then, it is not hard to see that (4.64) is equivalent to

$$\pi(\mathbf{x})\sum_{\mathbf{x}'\in\mathcal{X}} q(\mathbf{x},\mathbf{x}')=\sum_{\mathbf{x}'\in\mathcal{X}}\pi(\mathbf{x}')q(\mathbf{x}',\mathbf{x}), \qquad \mathbf{x}\in\mathcal{X}.$$

Thus, we can find the time-reversed process of $\{\mathbf{X}(t)\}$, which has the transition rate function:

$$\tilde{q}(\mathbf{x},\mathbf{x}')=\frac{\pi(\mathbf{x}')}{\pi(\mathbf{x})}q(\mathbf{x}',\mathbf{x}).$$

Then, it is not hard to see the following result.

Corollary 4.9. Under all the conditions of 4.8, the time-reversed macro process can be considered as that of the RGSMP with the speeds $\tilde{c}_{\mathbf{x}U}$ and the rates of the exponential distributions $\tilde{\mu}_{\gamma_{\mathbf{x}}(U)}$ and jump transition $\tilde{p}((\mathbf{x},U),(\mathbf{x}',\ell))$ as long as they satisfy

$$\tilde{c}_{\mathbf{x}U}\tilde{\mu}_{\gamma_{\mathbf{x}}(U)} = \frac{1}{\pi(\mathbf{x})} \sum_{\mathbf{x}' \in \mathcal{X}} \sum_{\ell \in J(\mathbf{x}')} c_{\mathbf{x}'\ell}\mu_{\gamma_{\mathbf{x}'}(\ell)}\pi(\mathbf{x}')p((\mathbf{x}',\ell),(\mathbf{x},U)), \tag{4.65}$$

$$\tilde{p}((\mathbf{x},U),(\mathbf{x}',\ell)) = \frac{1}{\pi(\mathbf{x})\tilde{c}_{\mathbf{x}U}\tilde{\mu}_{\gamma_{\mathbf{x}}(U)}} c_{\mathbf{x}'\ell}\mu_{\gamma_{\mathbf{x}'}(\ell)}\pi(\mathbf{x}')p((\mathbf{x}',\ell),(\mathbf{x},U)). \tag{4.66}$$

Remark 4.8. If U is not a singleton in this corollary, then clocks in U are forced to expire except for one in the constructed RGSMP for reversed time. However, active clocks are singly created. This is contrasted with the forward process.

Example 4.16 (Reversibility of the $M/M/1$ queue). We show how 4.9 can be used for applications. Consider the $M/M/1$ queue with arrival rate λ and service rate μ. We assume the stability condition $\rho \equiv \frac{\lambda}{\mu} < 1$. This model is a special case of the $M/G/1$ queue, which is formulated by the RGSMP in Example 4.15, and we can apply 4.9 because all lifetime distributions are exponential. Since $\hat{F}(\theta) = \mu/(\mu+\theta)$, (4.63) becomes

$$\hat{v}_*(z,0) = \frac{1-\rho}{1-\rho z}.$$

Hence, the stationary distribution $\{\pi(n)\}$ is given by $\pi(n) = (1-\rho)\rho^n$, as is well known. Remind that $D = D_e = \{0,1\}$, $J(0) = \{0\}$, and $J(n) = \{0,1\}$ for $n \geq 1$. From (4.65), we have, for $n \geq 0$,

$$\tilde{c}_{(n+1)0}\tilde{\mu}_0 = \frac{1}{\pi(n+1)}\lambda\pi(n) = \lambda\rho^{-1} = \mu,$$
$$\tilde{c}_{n1}\tilde{\mu}_1 = \frac{1}{\pi(n)}\mu\pi(n+1) = \mu\rho = \lambda,$$

Similarly we have

$$\tilde{p}((n,U),(n+1,1)) = p((n+1,1),(n,U)) = 1,$$
$$\tilde{p}((n+1,U),(n,0)) = p((n,0),(n+1,U)) = 1.$$

Thus, letting $\tilde{c}_{(n+1)0} = \tilde{c}_{n1} = 1$ for $n \geq 0$, the time reversed RGSMP is identical with the same $M/M/1$ queue except for the indexes, which are exchanged. Thus, the departure process of the original $M/G/1$ queue is the Poisson process with rate λ and independent of the past history of the system. This is known as the Burke's theorem [7], which is obtained for the $M/M/s$ queue. □

4.16 Product form decomposability

Under the assumptions of 4.8, $\{\mathbf{X}(t)\}$ is a continuous time Markov chain. However, this is not the case if there are non exponential lifetime distributions, so π determined by (4.64) may not be the stationary distribution of the macro state. We are interested in the case that this π is either still the stationary distribution of $\mathbf{X}(t)$ or can be modified to be the stationary distribution. We guess this could occurs when the remaining lifetimes are independent, and give the following definition.

Definition 4.13. If RGSMP $\{(\mathbf{X}(t),\mathbf{Y}(t))\}$ is stationary and if there exist distribution function H_d for each $d \in D$ such that $H_d(0) = 0$ and, under the stationary probability measure P,

$$P\Big(\mathbf{X}(0) = \mathbf{x}, \mathbf{Y}(0) \in \prod_{\ell=1}^{m(\mathbf{x})}[0,u_\ell]\Big) = P(\mathbf{X}(0) = \mathbf{x}) \prod_{\ell=1}^{m(\mathbf{x})} H_{\gamma_{\mathbf{x}}(\ell)}(u_\ell), \quad \mathbf{x} \in \mathcal{X}, u_\ell \geq 0,$$

then the RGSMP or its stationary distribution is said to have product form decomposition with respect to remaining lifetimes

Remark 4.9. The product form decomposability is slightly different from the conditionally independence of $Y_\ell(t)$ $(\ell = 1,2,\ldots,m(\mathbf{X}(t))$ given $\mathbf{X}(t) = \mathbf{x}$, that is,

$$P\Big(\mathbf{X}(t) = \mathbf{x}, \mathbf{Y}(t) \in \prod_{\ell=1}^{m(\mathbf{x})}[0,u_\ell]\Big) = P(\mathbf{X}(t) = \mathbf{x}) \prod_{\ell=1}^{m(\mathbf{x})} P(Y_\ell(t) \leq u_\ell).$$

Clearly, they are equivalent if no reallocation occurs.

Lemma 4.13. Assume that the RGSMP satisfies the assumptions (4.15a) and (4.15c). If the RGSMP has a product form decomposable stationary distribution ν, then (4.15b) is satisfied, and there exists $\alpha_d > 0$ for each $d \in D$ such that

$$H_d(x) = 1 - \beta_d \int_x^\infty (1 - F_d(u)) e^{-\alpha_d(u-x)} du, \qquad x \geq 0, d \in D, \tag{4.67}$$

$$c_{\mathbf{x}\ell}\mu^*_{\gamma_{\mathbf{x}}(\ell)} \hat{\nu}(\mathbf{x}, \theta_{\mathbf{x}}(\ell)) = \lambda_\ell \hat{\nu}_\ell(\mathbf{x}, \theta_{\mathbf{x}}(\ell)), \qquad \mathbf{x} \in \mathcal{X}, \ell \in J(\mathbf{x}), \tag{4.68}$$

where β_d and μ^*_d are given by

$$\beta_d = \begin{cases} \frac{\alpha_d}{1-\hat{F}_d(\alpha_d)}, & \alpha_d \neq 0, \\ \mu_d & \alpha_d = 0\,, \end{cases} \qquad \mu^*_d = \begin{cases} \frac{\alpha_d \hat{F}_d(\alpha_d)}{1-\hat{F}_d(\alpha_d)} & \alpha_d \neq 0, \\ \mu_d & \alpha_d = 0\,. \end{cases} \tag{4.69}$$

Proof. Assume that $(X(t),Y(t))$ is a stationary process with the stationary distribution ν. When F_d is exponential, we obviously have (4.67), and (4.68) is easily obtained similarly to (4.60). Hence, it is sufficient to prove (4.67) and (4.68) for $d \in D_{\rm g}$ and for $\ell \in J_{\rm g}(\mathbf{x})$. If $c_{\mathbf{x}\ell} = 0$, then (4.68) obviously holds, so we assume that $c_{\mathbf{x}\ell} > 0$. Let $T_{\ell 1} = \inf\{t > 0; N_\ell((0,t]) = 1\}$. Since $Y_\ell(0) = c_{\mathbf{x}\ell}T_{\ell 1}$, we have, from 4.1,

$$\begin{aligned} &\lambda_\ell P_\ell\left(\mathbf{X}(0-) = \mathbf{x}, Y_i(0-) \leq u_i, i \in J(\mathbf{x}) \setminus \{\ell\}\right) \\ &\quad = \lim_{t \downarrow 0} \frac{1}{t} P\left(\mathbf{X}(T_{\ell 1}-) = \mathbf{x}, Y_i(T_{\ell 1}-) \leq u_i, i \in J(\mathbf{x}) \setminus \{\ell\}, T_{\ell 1} \leq t\right) \\ &\quad = \lim_{t \downarrow 0} \frac{c_{\mathbf{x}\ell}}{c_{\mathbf{x}\ell}t} P\left(\mathbf{X}(0) = \mathbf{x}, Y_i(0) - c_{\mathbf{x}i}T_{\ell 1} \leq u_i, i \in J(\mathbf{x}) \setminus \{\ell\}, Y_\ell(0) \leq c_{\mathbf{x}\ell}t\right). \end{aligned}$$

Hence, from the product form decomposability and the definition of Palm distribution, we have

$$\begin{aligned}\lambda_\ell \nu_\ell \left(\mathbf{x}, \prod_{i=1}^{m(\mathbf{x})} [0, u_i] \right) &= c_{\mathbf{x}\ell} \frac{\partial}{\partial u_\ell} \nu \left(\mathbf{x}, \prod_{i=1}^{m(\mathbf{x})} [0, u_i] \right) \Bigg|_{u_\ell = 0} \\ &= c_{\mathbf{x}\ell} H'_{\gamma_{\mathbf{x}}(\ell)}(0) \pi(\mathbf{x}) \prod_{i \in J(\mathbf{x}) \setminus \{\ell\}} H_{\gamma_{\mathbf{x}}(i)}(u_i), \qquad (4.70)\end{aligned}$$

where $H'_{\gamma_{\mathbf{x}}(\ell)}(0)$ must exist and be finite because the left-hand side is finite. Note that this formula also holds for $\ell \in J_{\mathrm{e}}(\mathbf{x})$. Furthermore, if more than one clocks simultaneously expire, then we can put $u_i = u_\ell$ for some $i \neq \ell$ in the right-hand side of (4.70), which implies that the corresponding Palm distribution vanishes. Thus, (4.15b) is satisfied.

Letting $u_i = \infty$ in (4.70) and summing both sides of it for all $\mathbf{x} \in \mathcal{X}$ and $\ell \in \gamma_{\mathbf{x}}^{-1}(d)$ for each $d \in D$, we have

$$\sum_{\ell \in \gamma_{\mathbf{x}}^{-1}(d)} \lambda_\ell = H'_d(0) \sum_{\mathbf{x} \in \mathcal{X}} \sum_{\ell \in \gamma_{\mathbf{x}}^{-1}(d)} c_{\mathbf{x}\ell} \pi(\mathbf{x}). \qquad (4.71)$$

Thus, the right-hand side is finite by the assumption (4.15c).

Let $d = \gamma_{\mathbf{x}}(\ell)$, and letting $\theta_i = 0$ for all $i \in J(\mathbf{x}) \setminus \{\ell\}$ and $\theta_\ell = \theta$ in (4.61) of Theorem 4.10 and substituting (4.70) yield, for $\ell \in J_g(\mathbf{x})$,

$$\begin{aligned}c_{\mathbf{x}\ell} \theta \pi(\mathbf{x}) \hat{H}_d(\theta) &= c_{\mathbf{x}\ell} H'_d(0) \pi(\mathbf{x}) + \sum_{i \in J(\mathbf{x}) \setminus \{\ell\}} c_{\mathbf{x}i} H'_{\gamma_{\mathbf{x}}(i)}(0) \pi(\mathbf{x}) \hat{H}_d(\theta) \\ &- \sum_{\mathbf{x}' \in \mathcal{X}} \sum_{i \in J(\mathbf{x}')} \sum_{\ell \notin U \subset J(\mathbf{x})} c_{\mathbf{x}'i} H'_{\gamma_{\mathbf{x}'}(i)}(0) \pi(\mathbf{x}') p((\mathbf{x}', i), (\mathbf{x}, U)) \hat{H}_d(\theta) \\ &- \sum_{\mathbf{x}' \in \mathcal{X}} \sum_{i \in J(\mathbf{x}')} \sum_{\ell \in U \subset J(\mathbf{x})} c_{\mathbf{x}'i} H'_{\gamma_{\mathbf{x}'}(i)}(0) \pi(\mathbf{x}') p((\mathbf{x}', i), (\mathbf{x}, U)) \hat{F}_d(\theta). \qquad (4.72)\end{aligned}$$

Summing this formula for all $\mathbf{x} \in \mathcal{X}$ and $\ell \in \gamma_{\mathbf{x}}^{-1}(d)$ for each fixed $d \in D_g$, we can see that the sum of the left-hand side is finite by (4.71) and the second sum is not less than the third sum because of the interruption (they must be identical if there is no interruption). So, there exist a nonnegative constant a and positive constants b, c such that

$$\theta \hat{H}_d(\theta) = c + a\hat{H}_d(\theta) - b\hat{F}_d(\theta).$$

Letting $\theta = 0$ in this equation, we have $c = b - a$. Thus, we have, rewriting θ as θ,

$$\theta \hat{H}_d(\theta) - a\hat{H}_d(\theta) = -a + b(1 - \hat{F}_d(\theta)). \qquad (4.73)$$

This is equivalent to the following differential equation.

$$\frac{d}{dx} H_d(x) - aH_d(x) = -a + b(1 - F_d(x)) \qquad (4.74)$$

In fact, using the fact that $H_d(0) = 0$, one can check this equivalence by integrating both sides of the above equation multiplying $e^{-\theta x}$ concerning x over $[0,\infty)$ for each $\theta > 0$. We can easily solve the linear differential equation (4.74) using the boundary conditions $H_d(0) = 0$ and $\lim_{x\to\infty} H_d(x) = 1$. Thus, we get

$$H_d(x) = 1 - b\int_x^{\infty} (1 - F_d(x))e^{-a(u-x)}du, \qquad x \geq 0.$$

Denote a by α_d. Then, $b = \beta_b$, and we have (4.67). From (4.74), we have

$$H'_d(0) = \beta_d - \alpha_d = \mu^*_d.$$

Hence, (4.70) implies (4.68). □

Remark 4.10. From (4.73) and the expression of β_d, we have

$$\hat{H}_d(\theta) = \beta_d \frac{\hat{F}_d(\alpha_d) - \hat{F}_d(\theta)}{\theta - \alpha_d}, \qquad \theta \geq 0, d \in D_g. \tag{4.75}$$

From (4.68), we can interpret α_d as the rate for the interruption of a clock with index d. This rate does not depend on the macro state $\mathbf{x}$ and site $\ell \in J(\mathbf{x})$ as long as $d = \gamma_{\mathbf{x}}(\ell)$.

Lemma 4.14. Under the assumptions of 4.13, we have, for each $\mathbf{x} \in \mathcal{X}$,

$$\sum_{i\in J(\mathbf{x})} c_{\mathbf{x}i}\mu^*_{\gamma_{\mathbf{x}}(i)}\pi(\mathbf{x}) = \sum_{\mathbf{x}'\in\mathcal{X}}\sum_{i\in J(\mathbf{x}')}\sum_{U\subset J(\mathbf{x})} c_{\mathbf{x}'i}\mu^*_{\gamma_{\mathbf{x}'}(i)}\pi(\mathbf{x}')p((\mathbf{x}',i),(\mathbf{x},U)), \tag{4.76}$$

$$\begin{aligned} &c_{\mathbf{x}\ell}(\alpha_{\gamma_{\mathbf{x}}(\ell)} + \mu^*_{\gamma_{\mathbf{x}}(\ell)})\pi(\mathbf{x}) \\ &\quad = \sum_{\mathbf{x}'\in\mathcal{X}}\sum_{i\in J(\mathbf{x}')}\sum_{\ell\in U\subset J(\mathbf{x})} c_{\mathbf{x}'i}\mu^*_{\gamma_{\mathbf{x}'}(i)}\pi(\mathbf{x}')p((\mathbf{x}',i),(\mathbf{x},U)), \quad \ell \in J_g(\mathbf{x}). \end{aligned} \tag{4.77}$$

Remark 4.11. The right-hand side of (4.77) is the rate for the event that a new clock is activated at site ℓ. On the other hand, the left-hand side is the expiring rate of a clock at site ℓ. Hence, (4.77) represents the balance of the rates for expiring and activating clocks at the same site ℓ, so it is referred to as a local balance at site ℓ.

Proof. Substituting $H'_{\gamma_{\mathbf{x}'}(i)}(0) = \mu^*_{\gamma_{\mathbf{x}'}(i)}$ into (4.72) with $\theta = 0$ in the proof of 4.13 and noting the fact that (4.72) also holds for $\ell \in J_e(\mathbf{x})$, we have (4.76) for their summation. We next consider (4.72). For this let

$$\begin{aligned} K_1(\mathbf{x},\ell) = &\sum_{i\in J(\mathbf{x})\setminus\{\ell\}} c_{\mathbf{x}i}\mu^*_{\gamma_{\mathbf{x}}(i)}\pi(\mathbf{x}) \\ &- \sum_{\mathbf{x}'\in\mathcal{X}}\sum_{i\in J(\mathbf{x}')}\sum_{\ell\notin U\subset J(\mathbf{x})} c_{\mathbf{x}'i}\mu^*_{\gamma_{\mathbf{x}'}(i)}\pi(\mathbf{x}')p((\mathbf{x}',i),(\mathbf{x},U)), \end{aligned} \tag{4.78}$$

$$K_2(\mathbf{x},\ell) = \sum_{\mathbf{x}'\in\mathcal{X}}\sum_{i\in J(\mathbf{x}')}\sum_{\ell\in U\subset J(\mathbf{x})} c_{\mathbf{x}'i}\mu^*_{\gamma_{\mathbf{x}'}(i)}\pi(\mathbf{x}')p((\mathbf{x}',i),(\mathbf{x},U)). \tag{4.79}$$

Then, (4.72) can be written as

$$c_{\mathbf{x}\ell}\theta_\ell\pi(\mathbf{x})\hat{H}_d(\theta) = c_{\mathbf{x}\ell}H'_d(0)\pi(\mathbf{x}) + K_1(\mathbf{x},\ell)\hat{H}_d(\theta) - K_2(\mathbf{x},\ell)\hat{F}_d(\theta)$$

Subtracting this from (4.73) multiplied by $c_{\mathbf{x}\ell}\theta_\ell\pi(\mathbf{x})$, we have

$$(\alpha_d c_{\mathbf{x}\ell}\pi(\mathbf{x}) - K_1(\mathbf{x},\ell))(1-\hat{H}_d(\theta)) = (\beta_d c_{\mathbf{x}\ell}\pi(\mathbf{x}) - K_2(\mathbf{x},\ell))(1-\hat{F}_d(\theta)).$$

Because $1-\hat{H}_d(\theta)$ can not be a constant multiplication of $(1-\hat{F}_d(\theta))$ for $d \in D_g$ by 4.13, their coefficients must vanish. Thus, we have

$$\alpha_d c_{\mathbf{x}\ell}\pi(\mathbf{x}) = K_1(\mathbf{x},\ell), \qquad \beta_d c_{\mathbf{x}\ell}\pi(\mathbf{x}) = K_2(\mathbf{x},\ell). \tag{4.80}$$

This is nothing but (4.77) because $\beta_d = \alpha_d + \mu_d^*$. □

It is notable that (4.76) represents the global balance under macro state $\mathbf{x}$ while (4.77) is the local balance at site ℓ under macro state $\mathbf{x}$. We are now ready to prove the following theorem.

Theorem 4.11. The RGSMP satisfying the assumptions (4.15a) is product form decomposable and satisfies and (4.15c) if and only if there exist the distribution π on $\mathcal{X}$ and nonnegative numbers $\{\alpha_d; d \in D_g\}$ satisfying the global balance (4.76), the local balance (4.77) and the finite intensity condition:

$$\sum_{\mathbf{x}\in\mathcal{X}}\sum_{\ell\in J(\mathbf{x})} c_{\mathbf{x}\ell}\mu^*_{\gamma_{\mathbf{x}}(\ell)}\pi(\mathbf{x}) < \infty. \tag{4.81}$$

In this case, the stationary distribution ν is given by

$$\nu\Big(\mathbf{x}, \prod_{i\in J(\mathbf{x})}[0,u_i]\Big) = \pi(\mathbf{x})\prod_{i\in J(\mathbf{x})} H_{\gamma_{\mathbf{x}}(i)}(u_i), \quad \mathbf{x}\in\mathcal{X}, u_i \ge 0, \tag{4.82}$$

where $\alpha_d = 0$ for $d \in D_e$ and μ_d^* and H_d are defined in 4.13. Furthermore, under this stationary distribution, (4.15b) is satisfied, and not more than one clock is activated at once, that is, we have, for each $\mathbf{x}\in\mathcal{X}$ and any $\ell_1,\ell_2 \in J_g(\mathbf{x})$ such that $\ell_1 \neq \ell_2$,

$$\sum_{\mathbf{x}'\in\mathcal{X}}\sum_{i\in J(\mathbf{x}')}\sum_{\{\ell_1,\ell_2\}\subset U\subset J_g(\mathbf{x})} c_{\mathbf{x}'i}\mu^*_{\gamma_{\mathbf{x}'}(i)}\pi(\mathbf{x}')p((\mathbf{x}',i),(\mathbf{x},U)) = 0. \tag{4.83}$$

Proof. We have already shown that the product decomposability with conditions (4.15a), (4.15b) and (4.15c) implies (4.76), (4.77), (4.81) and (4.82) with the nonnegative number α_d for $d \in D$. Thus, for the necessity, we only need to prove the last statement. Suppose that $\ell_1,\ell_2 \in J_g(\mathbf{x})$ satisfying $\ell_1 \neq \ell_2$ are simultaneously activated. Let $d_i = \gamma_{\mathbf{x}}(\ell_i)$ for $i = 1,2$. Then, similar to (4.72), it follows from (4.61) that

$$
\begin{aligned}
&(c_{\mathbf{x}\ell_1}\theta_1 + c_{\mathbf{x}\ell_2}\theta_2)\pi(\mathbf{x})\hat{H}_{d_1}(\theta_1)\hat{H}_{d_2}(\theta_2)\\
&\quad = (c_{\mathbf{x}\ell_1}\mu^*_{d_1}\hat{H}_{d_1}(\theta_1) + c_{\mathbf{x}\ell_2}\mu^*_{d_2}\hat{H}_{d_2}(\theta_2))\pi(\mathbf{x})\\
&\qquad + \sum_{i\in J(\mathbf{x})\setminus\{\ell_1,\ell_2\}} c_{\mathbf{x}i}\mu^*_{\gamma_{\mathbf{x}}(i)}\pi(\mathbf{x})\hat{H}_{d_1}(\theta_1)\hat{H}_{d_2}(\theta_2)\\
&\qquad - \sum_{\mathbf{x}'\in\mathcal{X}}\sum_{i\in J(\mathbf{x}')}\sum_{\ell_1,\ell_2\notin U\subset J(\mathbf{x})} c_{\mathbf{x}'i}\mu^*_{\gamma_{\mathbf{x}'}(i)}\pi(\mathbf{x}')p((\mathbf{x}',i),(\mathbf{x},U))\hat{H}_{d_1}(\theta_1)\hat{H}_{d_2}(\theta_2)\\
&\qquad - \sum_{\mathbf{x}'\in\mathcal{X}}\sum_{i\in J(\mathbf{x}')}\sum_{\ell_1\notin U,\ell_2\in U\subset J(\mathbf{x})} c_{\mathbf{x}'i}\mu^*_{\gamma_{\mathbf{x}'}(i)}\pi(\mathbf{x}')p((\mathbf{x}',i),(\mathbf{x},U))\hat{H}_{d_1}(\theta_1)\hat{F}_{d_2}(\theta_2)\\
&\qquad - \sum_{\mathbf{x}'\in\mathcal{X}}\sum_{i\in J(\mathbf{x}')}\sum_{\ell_1\in U,\ell_2\notin U\subset J(\mathbf{x})} c_{\mathbf{x}'i}\mu^*_{\gamma_{\mathbf{x}'}(i)}\pi(\mathbf{x}')p((\mathbf{x}',i),(\mathbf{x},U))\hat{F}_{d_1}(\theta_1)\hat{H}_{d_2}(\theta_2)\\
&\qquad - \sum_{\mathbf{x}'\in\mathcal{X}}\sum_{i\in J(\mathbf{x}')}\sum_{\ell_1,\ell_2\in U\subset J(\mathbf{x})} c_{\mathbf{x}'i}\mu^*_{\gamma_{\mathbf{x}'}(i)}\pi(\mathbf{x}')p((\mathbf{x}',i),(\mathbf{x},U))\hat{F}_{d_1}(\theta_1)\hat{F}_{d_2}(\theta_2). \quad (4.84)
\end{aligned}
$$

On the other hand, multiplying (4.72) for $\ell = \ell_1$ and $\theta = \theta_1$ by $\hat{H}_{d_2}(\theta_2)$, we have

$$
\begin{aligned}
&c_{\mathbf{x}\ell_1}\theta_1\pi(\mathbf{x})\hat{H}_{d_1}(\theta_1)\hat{H}_{d_2}(\theta_2)\\
&\quad = (c_{\mathbf{x}\ell_1}\mu^*_{d_1}\hat{H}_{d_1}(\theta_1)\hat{H}_{d_2}(\theta_2) + c_{\mathbf{x}\ell_2}\mu^*_{d_2}\hat{H}_{d_2}(\theta_2))\pi(\mathbf{x})\\
&\qquad + \sum_{i\in J(\mathbf{x})\setminus\{\ell_1,\ell_2\}} c_{\mathbf{x}i}\mu^*_{\gamma_{\mathbf{x}}(i)}\pi(\mathbf{x})\hat{H}_{d_1}(\theta_1)\hat{H}_{d_2}(\theta_2)\\
&\qquad - \sum_{\mathbf{x}'\in\mathcal{X}}\sum_{i\in J(\mathbf{x}')}\sum_{\ell_1,\ell_2\notin U\subset J(\mathbf{x})} c_{\mathbf{x}'i}\mu^*_{\gamma_{\mathbf{x}'}(i)}\pi(\mathbf{x}')p((\mathbf{x}',i),(\mathbf{x},U))\hat{H}_{d_1}(\theta_1)\hat{H}_{d_2}(\theta_2)\\
&\qquad - \sum_{\mathbf{x}'\in\mathcal{X}}\sum_{i\in J(\mathbf{x}')}\sum_{\ell_1\notin U,\ell_2\in U\subset J(\mathbf{x})} c_{\mathbf{x}'i}\mu^*_{\gamma_{\mathbf{x}'}(i)}\pi(\mathbf{x}')p((\mathbf{x}',i),(\mathbf{x},U))\hat{H}_{d_1}(\theta_1)\hat{H}_{d_2}(\theta_2)\\
&\qquad - \sum_{\mathbf{x}'\in\mathcal{X}}\sum_{i\in J(\mathbf{x}')}\sum_{\ell_1\in U,\ell_2\notin U\subset J(\mathbf{x})} c_{\mathbf{x}'i}\mu^*_{\gamma_{\mathbf{x}'}(i)}\pi(\mathbf{x}')p((\mathbf{x}',i),(\mathbf{x},U))\hat{F}_{d_1}(\theta_1)\hat{H}_{d_2}(\theta_2)\\
&\qquad - \sum_{\mathbf{x}'\in\mathcal{X}}\sum_{i\in J(\mathbf{x}')}\sum_{\ell_1,\ell_2\in U\subset J(\mathbf{x})} c_{\mathbf{x}'i}\mu^*_{\gamma_{\mathbf{x}'}(i)}\pi(\mathbf{x}')p((\mathbf{x}',i),(\mathbf{x},U))\hat{F}_{d_1}(\theta_1)\hat{H}_{d_2}(\theta_2). \quad (4.85)
\end{aligned}
$$

Subtracting both sides (4.85) from (4.84), we have

$$
\begin{aligned}
&c_{\mathbf{x}\ell_2}\theta_2\pi(\mathbf{x})\hat{H}_{d_1}(\theta_1)\hat{H}_{d_2}(\theta_2) = c_{\mathbf{x}\ell_2}\mu^*_{d_2}\pi(\mathbf{x})(1-\hat{H}_{d_2}(\theta_2))\hat{H}_{d_1}(\theta_1)\\
&- \sum_{\mathbf{x}'\in\mathcal{X}}\sum_{i\in J(\mathbf{x}')}\sum_{\ell_1\notin U,\ell_2\in U\subset J(\mathbf{x})} c_{\mathbf{x}'i}\mu^*_{\gamma_{\mathbf{x}'}(i)}\pi(\mathbf{x}')p((\mathbf{x}',i),(\mathbf{x},U))\hat{H}_{d_1}(\theta_1))(\hat{F}_{d_2}(\theta_2)-\hat{H}_{d_2}(\theta_2))\\
&- \sum_{\mathbf{x}'\in\mathcal{X}}\sum_{i\in J(\mathbf{x}')}\sum_{\ell_1,\ell_2\in U\subset J(\mathbf{x})} c_{\mathbf{x}'i}\mu^*_{\gamma_{\mathbf{x}'}(i)}\pi(\mathbf{x}')p((\mathbf{x}',i),(\mathbf{x},U))\hat{F}_{d_1}(\theta_1)(\hat{F}_{d_2}(\theta_2)-\hat{H}_{d_2}(\theta_2)).
\end{aligned}
$$

Dividing both sides of the above formula by θ_2 and letting $\theta_2 \downarrow 0$ yield

$$
\begin{aligned}
&c_{\mathbf{x}\ell_2}\pi(\mathbf{x})\hat{H}_{d_1}(\theta_1) = c_{\mathbf{x}\ell_2}\mu^*_{d_2}\pi(\mathbf{x})(-\hat{H}'_{d_2}(0))\hat{H}_{d_1}(\theta_1)\\
&- \sum_{\mathbf{x}'\in\mathcal{X}}\sum_{i\in J(\mathbf{x}')}\sum_{\ell_1\notin U,\ell_2\in U\subset J(\mathbf{x})} c_{\mathbf{x}'i}\mu^*_{\gamma_{\mathbf{x}'}(i)}\pi(\mathbf{x}')p((\mathbf{x}',i),(\mathbf{x},U))\hat{H}_{d_1}(\theta_1))(\hat{F}'_{d_2}(0)-H'_{d_2}(0))\\
&- \sum_{\mathbf{x}'\in\mathcal{X}}\sum_{i\in J(\mathbf{x}')}\sum_{\ell_1,\ell_2\in U\subset J(\mathbf{x})} c_{\mathbf{x}'i}\mu^*_{\gamma_{\mathbf{x}'}(i)}\pi(\mathbf{x}')p((\mathbf{x}',i),(\mathbf{x},U))\hat{F}_{d_1}(\theta_1)(\hat{F}'_{d_2}(0)-H'_{d_2}(0)).
\end{aligned}
$$

Consequently, $\hat{F}_{d_1}(\theta_1)$ must be proportional to $\hat{H}_{d_1}(\theta_1)$ and therefore identical with $\hat{H}_{d_1}(\theta_1)$. This is impossible. Thus, not more than one clock can not be activated at once.

We next show the converse. Summing (4.77) over all $\ell \in J_g(\mathbf{x})$ and subtracting this sum from (4.76), we have

$$\begin{aligned}\sum_{i\in J_e(\mathbf{x})} c_{\mathbf{x}i}\mu^*_{\gamma_{\mathbf{x}}(i)}\pi(\mathbf{x}) &= \sum_{i\in J_g(\mathbf{x})} c_{\mathbf{x}i}\alpha_{\gamma_{\mathbf{x}}(i)}\pi(\mathbf{x}) \\ &\quad + \sum_{\mathbf{x}'\in\mathcal{X}}\sum_{i\in J(\mathbf{x}')}\sum_{U\subset J_e(\mathbf{x})} c_{\mathbf{x}'i}\mu^*_{\gamma_{\mathbf{x}'}(i)}\pi(\mathbf{x}')p((\mathbf{x}',i),(\mathbf{x},U)). \qquad (4.86)\end{aligned}$$

From the definition of $\hat{H}_d(\theta)$, it follows that

$$\beta_d\hat{F}_d(\theta) = (\alpha_d-\theta)\hat{H}_d(\theta)+\mu^*_d.$$

Multiplying both sides of (4.77) by $\hat{F}_d(\theta)$ and substituting the above $\hat{F}_d(\theta)$ to its left side, we have

$$c_{\mathbf{x}\ell}\theta\pi(\mathbf{x})\hat{H}_d(\theta) = c_{\mathbf{x}\ell}\mu^*_d\pi(\mathbf{x}) + c_{\mathbf{x}\ell}\alpha_d\pi(\mathbf{x})\hat{H}_d(\theta) - K_2(\mathbf{x},\ell)\hat{F}_d(\theta). \qquad (4.87)$$

From (4.76) and (4.77), we have

$$\begin{aligned}\sum_{i\in J(\mathbf{x})} c_{\mathbf{x}i}\mu^*_{\gamma_{\mathbf{x}}(i)}\pi(\mathbf{x}) &= \sum_{\mathbf{x}'\in\mathcal{X}}\sum_{i\in J(\mathbf{x}')}\sum_{\ell\notin U\subset J(\mathbf{x})} c_{\mathbf{x}'i}\mu^*_{\gamma_{\mathbf{x}'}(i)}\pi(\mathbf{x}')p((\mathbf{x}',i),(\mathbf{x},U)) \\ &\quad + \sum_{\mathbf{x}'\in\mathcal{X}}\sum_{i\in J(\mathbf{x}')}\sum_{\ell\in U\subset J(\mathbf{x})} c_{\mathbf{x}'i}\mu^*_{\gamma_{\mathbf{x}'}(i)}\pi(\mathbf{x}')p((\mathbf{x}',i),(\mathbf{x},U)) \\ &= \sum_{\mathbf{x}'\in\mathcal{X}}\sum_{i\in J(\mathbf{x}')}\sum_{\ell\notin U\subset J(\mathbf{x})} c_{\mathbf{x}'i}\mu^*_{\gamma_{\mathbf{x}'}(i)}\pi(\mathbf{x}')p((\mathbf{x}',i),(\mathbf{x},U)) \\ &\quad + c_{\mathbf{x}\ell}(\alpha_{\gamma_{\mathbf{x}}(\ell)}+\mu^*_{\gamma_{\mathbf{x}}(\ell)})\pi(\mathbf{x}).\end{aligned}$$

Substituting $c_{\mathbf{x}\ell}\alpha_{\gamma_{\mathbf{x}}(\ell)}\pi(\mathbf{x})$ from this equation into (4.87), we arrive at

$$\begin{aligned}c_{\mathbf{x}\ell}\theta\pi(\mathbf{x})\hat{H}_d(\theta) &= c_{\mathbf{x}\ell}\mu^*_d\pi(\mathbf{x}) + \sum_{i\in J(\mathbf{x})\setminus\{\ell\}} c_{\mathbf{x}i}\mu^*_{\gamma_{\mathbf{x}}(i)}\pi(\mathbf{x})\hat{H}_d(\theta) \\ &\quad - \sum_{\mathbf{x}'\in\mathcal{X}}\sum_{i\in J(\mathbf{x}')}\sum_{\ell\notin U\subset J(\mathbf{x})} c_{\mathbf{x}'i}\mu^*_{\gamma_{\mathbf{x}'}(i)}\pi(\mathbf{x}')p((\mathbf{x}',i),(\mathbf{x},U))\hat{H}_d(\theta) \\ &\quad - \sum_{\mathbf{x}'\in\mathcal{X}}\sum_{i\in J(\mathbf{x}')}\sum_{\ell\in U\subset J(\mathbf{x})} c_{\mathbf{x}'i}\mu^*_{\gamma_{\mathbf{x}'}(i)}\pi(\mathbf{x}')p((\mathbf{x}',i),(\mathbf{x},U))\hat{F}_d(\theta).\end{aligned}$$

This equation is identical with (4.72). Let $d=\gamma_{\mathbf{x}}(\ell)$, multiply both sides of it by $\prod_{i\in J(\mathbf{x})\setminus\{\ell\}}\hat{H}_{\gamma_{\mathbf{x}}(i)}(\theta_i)$ and define distributions ν,ν_ℓ and constants λ_ℓ by

$$\hat{\nu}(\mathbf{x}, \theta_{\mathbf{x}}(J_{\mathrm{e}}(\mathbf{x}))) = \pi(\mathbf{x}) \prod_{i \in J(\mathbf{x})} \hat{H}_{\gamma_{\mathbf{x}}(i)}(\theta_i),$$
$$\hat{\nu}_\ell(\mathbf{x}, \theta_{\mathbf{x}}(J_{\mathrm{e}}(\mathbf{x}))) = \pi(\mathbf{x}) \prod_{i \in J(\mathbf{x}) \setminus \{\ell\}} \hat{H}_{\gamma_{\mathbf{x}}(i)}(\theta_i),$$
$$\lambda_\ell = \sum_{\mathbf{x} \in \mathcal{X}} c_{\mathbf{x}\ell} \mu^*_{\gamma_{\mathbf{x}}(\ell)} \pi(\mathbf{x}).$$

We then have the stationary equation (4.61). Hence, ν is the stationary distribution of the RGSMP by Theorem 4.10. □

There are a number of remarks on this theorem.

Remark 4.12. This theorem does not answer the uniqueness of the stationary distribution. However, the uniqueness can be considered through the irreducibility. In particular, for the macro state distribution, it is not hard to check the irreducibility from the global balance equation (4.76) similar to the irreducibility of a Markov chain with discrete state space $\mathcal{X}$.

Remark 4.13. Although at most one clock with non exponentially distributed lifetime is activated at each completion time, some clocks with exponentially distributed life times may be activated at the same instant. Thus, it is not necessary that $U = \{\ell\}$ in (4.76) and (4.77).

Remark 4.14. From the proof of 4.14, we can see that $\alpha_{\gamma_{\mathbf{x}}(\ell)} > 0$ if and only if $K_1(\mathbf{x}, \ell) > 0$, that is

$$\sum_{i \in J(\mathbf{x}) \setminus \{\ell\}} c_{\mathbf{x}i} \mu^*_{\gamma_{\mathbf{x}}(i)} \pi(\mathbf{x}) - \sum_{\mathbf{x}' \in \mathcal{X}} \sum_{i \in J(\mathbf{x}')} \sum_{\ell \notin U \subset J(\mathbf{x})} c_{\mathbf{x}'i} \mu^*_{\gamma_{\mathbf{x}'}(i)} \pi(\mathbf{x}') p((\mathbf{x}', i), (\mathbf{x}, U)) > 0.$$

By the global equation (4.76), this is equivalent to

$$\sum_{\mathbf{x}' \in \mathcal{X}} \sum_{i \in J(\mathbf{x}')} \sum_{\ell \in U \subset J(\mathbf{x})} c_{\mathbf{x}'i} \mu^*_{\gamma_{\mathbf{x}'}(i)} \pi(\mathbf{x}') p((\mathbf{x}', i), (\mathbf{x}, U)) - c_{\mathbf{x}\ell} \mu^*_{\gamma_{\mathbf{x}}(\ell)} \pi(\mathbf{x}) > 0.$$

This means that $\alpha_{\gamma_{\mathbf{x}}(\ell)} > 0$ holds if and only if the total activation rate of type $d = \gamma_{\mathbf{x}}(\ell)$ clock is greater than its total completion rate. Thus, α_d can be interpreted as an interruption rate.

Remark 4.15. We have not discussed how to compute the interruption rate α_d. In many cases, they are given as modeling parameters. If this is not the case, they would be determined by (4.80) although they are highly nonlinear equations.

In the rest of this section, we consider the case where there is no interruption, that is, $\alpha_d = 0$ for all $d \in D$. The following corollary is immediate from Theorem 4.11.

Corollary 4.10. Suppose the RGSMP satisfies (4.15a) and has no interruption. Then, the RGSMP is product form decomposable and satisfies (4.15b) and (4.15c) if and only if there exist the distribution π on $\mathcal{X}$ satisfying the global and local balances:

$$\sum_{i\in J(\mathbf{x})} c_{\mathbf{x}i}\mu_{\gamma_{\mathbf{x}}(i)}\pi(\mathbf{x}) = \sum_{\mathbf{x}'\in\mathcal{X}}\sum_{i\in J(\mathbf{x}')}\sum_{U\subset J(\mathbf{x})} c_{\mathbf{x}'i}\mu_{\gamma_{\mathbf{x}'}(i)}\pi(\mathbf{x}')p((\mathbf{x}',i),(\mathbf{x},U)), \tag{4.88}$$

$$c_{\mathbf{x}\ell}\mu_{\gamma_{\mathbf{x}}(\ell)}\pi(\mathbf{x}) = \sum_{\mathbf{x}'\in\mathcal{X}}\sum_{i\in J(\mathbf{x}')}\sum_{\ell\in U\subset J(\mathbf{x})} c_{\mathbf{x}'i}\mu_{\gamma_{\mathbf{x}'}(i)}\pi(\mathbf{x}')p((\mathbf{x}',i),(\mathbf{x},U)),\ \ell\in J_g(\mathbf{x}), \tag{4.89}$$

and the finite intensity condition:

$$\sum_{\mathbf{x}\in\mathcal{X}}\sum_{\ell\in J(\mathbf{x})} c_{\mathbf{x}\ell}\mu_{\gamma_{\mathbf{x}}(\ell)}\pi(\mathbf{x}) < \infty. \tag{4.90}$$

In this case, the stationary distribution ν is given by

$$\nu\Big(\mathbf{x}, \prod_{i\in J(\mathbf{x})}[0,u_i]\Big) = \pi(\mathbf{x})\prod_{i\in J(\mathbf{x})}\mu_{\gamma_{\mathbf{x}}(i)}\int_0^{u_i}(1-F_{\gamma_{\mathbf{x}}(i)}(v))dv, \quad \mathbf{x}\in\mathcal{X}, u_i\geq 0. \tag{4.91}$$

Furthermore, not more than one clock is activated at once.

Note that the stationary distribution π of the macro states depend on F_d for $d\in D_g$ only through their means μ_d^{-1}. This stationary distribution is said to be insensitive with respect to F_d for $d\in D_g$.

Example 4.17. Consider the $M/G/1$ queue of Example 4.15. We have formulated it by the RGSMP. Since there is no interruption, we examine the product form decomposability by 4.10. The local balance condition (4.89) is

$$\mu_1\pi(1) = \lambda\pi(0) + \mu_1\pi(2), \qquad \mu_1\pi(n) = \mu_1\pi(n+1), \qquad n\geq 2.$$

Obviously, these are impossible. Hence, the $M/G/1$ queue with the first-come first-served discipline can not be product form decomposable. □

4.17 Applications to queues and their networks

How we can check the conditions in Theorem 4.11 and 4.10 to see the decomposability ? It is notable that we do not need to consider the RGSMP with generally distributed lifetimes. Namely, we only need to find the stationary distribution of the macro state which satisfies (4.76) and (4.77) (or (4.88) and (4.89)). In particular, if there is no interruption, it is sufficient to consider the RGSMP all of whose lifetimes are exponentially distributed. This greatly simplifies the verification of the decomposability.

In this section, we exemplify queues and their networks by applying 4.10 and Theorem 4.11 in this way. We first consider the following queueing system.

(4.17a) There are service positions numbered $1,2,\ldots$ to accommodate one customer in each position. Customers arrives subject to the Poisson process with rate λ with *i.i.d* amounts of work for service, whose distribution is denoted by F. This F is assumed to have a finite mean $\frac{1}{\mu}$.

(4.17b) An arriving customer who found n customers in the system gets into position ℓ with probability $\delta_{n+1,\ell}$ for $\ell = 1, 2, \cdots, n+1$, and customers in positions $\ell, \ell+1, \cdots, n$ move to $\ell+1, \ell+2, \cdots, n+1$, respectively, where

$$\sum_{\ell=1}^{n+1} \delta_{n+1,\ell} = 1, \qquad n \geq 0.$$

Thus, if there are n customers in the system, positions $1, 2, \cdots, n$ are occupied.

(4.17c) A customer in position ℓ is served at rate $c_{n,\ell}$ for $\ell = 1, 2, \ldots, n$ when there are n customers in the system. Denote the total service rate in this case by $\sigma(n)$. That is,

$$\sigma(n) = \sum_{\ell=1}^{n} c_{n\ell}, \qquad n \geq 1.$$

If a customer in position ℓ leaves the system, customers in positions $\ell+1, \ell+2, \cdots, n$ move to $\ell, \ell+1, \cdots, n-1$.

This model is referred to as a packed positioning queue. For each t, denote the number of customers in system by $X(t)$ and the remaining work of the customer at position ℓ by $Y_\ell(t)$ for $\ell = 1, 2, \ldots, X(t)$. Let $\mathbf{Y}(t) = (Y_1(t), \ldots, Y_{X(t)}(t))$. We show that the process $(X(t), \mathbf{Y}(t))$ is the RGSMP. Let

$$\mathcal{X} = \mathbb{N}_+, \quad D = \{e, g\}, \quad J_e(n) = \{0\}, \quad J_g(n) = \{1, 2, \ldots, n\} \text{ for } n \in \mathcal{X},$$

where $\mathbb{N}_+ = \{0, 1, 2, \ldots\}$. The index functions are defined as $\gamma_{en}(0) = e$ and $\gamma_{gn}(\ell) = g$ for $\ell \geq 1$, where e and g represent exponential and general distributions, respectively.

Let $c_{n0} = \lambda$ for all $n \in \mathcal{X}$. We interpret $c_{n\ell}$ for $\ell \geq 1$ as the speed of a clock at the site ℓ under the macrostate n. Define transition probabilities by

$$\begin{aligned} &p((n,0),(n+1,\ell)) = \delta_{(n+1)\ell}, && (n \in \mathcal{X}, 1 \leq \ell \leq n+1), \\ &p((n,\ell),(n-1,\emptyset)) = 1, && (n \geq 1, 1 \leq \ell \leq n), \\ &\mu_e = \lambda, \qquad \mu_g = \mu. \end{aligned}$$

Thus, $(X(t), \mathbf{Y}(t))$ can be considered as the RGSMP. Hence, by 4.10, the RGSMP supplemented by the remaining service requirements is product-form decomposable if and only if

$$c_{n\ell} \mu \pi(n) = \lambda \delta_{n\ell} \pi(n-1) \qquad (n \geq 1, 1 \leq \ell \leq n). \tag{4.92}$$

From this, we see that

$$c_{n\ell} = \sigma(n) \delta_{n\ell}, \qquad n \geq 1, \ell = 1, 2, \ldots, n.$$

This service discipline is called symmetric by Kelly [18, 19].

Thus, the packed positioning queue with Poisson arrivals and *i.i.d.* service requirements is product form decomposable if and only if its service discipline is symmetric. In this case, (4.92) uniquely determines the stationary distribution $\{\pi(n)\}$ as

$$\pi(n) = \pi(0)\frac{\lambda^n}{\mu^n \prod_{i=1}^n \sigma(i)}, \qquad n \geq 1.$$

where $\sigma(n) = \sum_{\ell=1}^n c_{n\ell}$, if

$$\sum_{n=0}^{+\infty} \frac{\lambda^n}{\mu^n \prod_{i=1}^n \sigma(i)} < +\infty.$$

This stationary distribution is insensitive with respect to the work for service. Furthermore, (4.92) implies

$$\sum_{\ell=1}^{n+1} c_{n\ell}\mu\pi(n+1) = \lambda\pi(n) \qquad n \geq 0.$$

Hence, by a similar time-reversed argument in Example 4.16, we can see that the departure process from this queue is also Poisson. This is generally known as quasi-reversibility (see [9] for its details).

Note that packing rule of service positions does not affect to get (4.92), that is, any reallocations are possible at arriving and departing instants if all positions are packed. In what follows, we refer to this model simply as a symmetric queue.

Remark 4.16. One might expect that the insensitivity of the queue length distribution implies the symmetric condition. But, this is not true. For example, assume that, for each n, $c_{n\ell} = \frac{1}{n}$ for $1 \leq \ell \leq n$ and $\delta_{n\ell} = 1$ only if $\ell = 1$. Then, the sample path of $\{X(t)\}$ is identical with that of the corresponding symmetric queue with $c_{n\ell} = \delta_{n\ell} = \frac{1}{n}$ for $1 \leq \ell \leq n$ since all customers in service have a same service rate after arriving of a new customer. Thus, (4.92) does not hold but the queue length distribution is still insensitive. This example is rather trivial, but shows the local balance (4.76) is indeed stronger than the insensitivity. □

We next consider the case where interruptions occur in the symmetric queue with Poisson arrivals. In addition to the assumptions (4.17a), (4.17b) and (4.17c), we assume the following condition.

(4.17d) Negative signals arrive according to the Poisson process with rate $\alpha\sigma(n)$, which is independent of everything else, and delete a customer in position ℓ with probability $\delta_{n\ell}$ when n customers are in the system.

The index for this signal is denoted by -1. That is, $J(n) = \{-1, 0, 1, 2, \ldots, n\}$ for $n \geq 0$. Suppose the local balance (4.77) holds. Since the general index g is only activated by arrivals, we have

$$(\mu^* + \alpha)\delta_{n\ell}\sigma(n)\pi(n) = \lambda\delta_{n\ell}\pi(n-1), \qquad n \geq 1,\ \ell = 1, 2, \ldots, n,$$

where μ^* is given by (4.69) for $\alpha_d = \alpha$. Thus, the stationary distribution is given by

$$\pi(n) = \pi(0)\frac{\lambda^n}{(\mu^* + \alpha)^n \prod_{i=1}^n \sigma(i)}, \qquad n \geq 0,$$

where the stability condition $\sum_{n=0}^{\infty} \frac{\lambda^n}{(\mu^* + \alpha)^n \prod_{i=1}^n \sigma(i)} < \infty$ is assumed. Then, it is easy to see that this distribution satisfies the global balance (4.76):

$$(\lambda + (\mu^* + \alpha)\sigma(n))\pi(n) = \lambda\pi(n-1) + (\mu^* + \alpha)d(n+1)\pi(n+1), \quad n \geq 1.$$

Hence, (4.77) indeed holds, and we have the product form decomposability by Theorem 4.11.

Example 4.18. The symmetric queue can be generalized for multi-class queues and their networks. We show how to formulate multi-class symmetric queues by an RGSMP. Suppose there are T types of customers. Denote a set of their types $\{1, 2, \cdots, T\}$ by $\mathcal{T}$. We assume that the arrival process of type i customers is Poisson with the rate λ_i and the arrival streams of different types of customers are independent. Now the macrostate needs to specify the configuration of customer types in positions. So far, we let

$$\mathcal{X} = \{\mathbf{x} = (t(1), t(2), \cdots, t(n)); n \geq 0, t(i) \in \mathcal{T}\}.$$

The site space is same as the packed positioning queue, but D is changed to $\{e, 1, \cdots, T\}$, which means that different types of customers may have different service time distributions. Service discipline is also same as the packed positioning queue. We assume that the speeds of service and position selecting probabilities of arriving customers may depends on $n = |\mathbf{x}|$, so the total speed also only depends on n, which is denoted by $\sigma(n)$. Then, the local balance (4.76) becomes, for $\mathbf{x} = (t(1), \cdots, t(n))$ and $\mathbf{x} \ominus \mathbf{e}_\ell = (t(1), \cdots, t(\ell-1), t(\ell+1), \cdots, t(n))$,

$$c_{n\ell}\mu_{t(\ell)}\pi(\mathbf{x}) = \delta_{n\ell}\lambda_{t(\ell)}\pi(\mathbf{x} \ominus \mathbf{e}_\ell), \qquad \ell \in J_g(\mathbf{x}) \equiv \{1, 2, \ldots, n\}. \tag{4.93}$$

Thus, if the queue is symmetric, i.e., if $c_{n\ell}$ is proportional to $\delta_{n\ell}$ concerning ℓ for each $n = |\mathbf{x}|$, then

$$\pi(\mathbf{x}) = \pi(\mathbf{0})\prod_{\ell=1}^{n} \frac{\lambda_{t(\ell)}}{\mu_{t(\ell)}\sigma(\ell)}, \qquad \mathbf{x} = (t(1), \cdots, t(n)),$$

gives a stationary distribution if the total sum of $\pi(\mathbf{x})$ over $\mathcal{X}$ is finite, where $\pi(\mathbf{0})$ is the normalizing constant. From (4.93), we again have the quasi-reversibility for each fixed type t as

$$\sum_{\ell=1}^{n+1} c_{n\ell}\mu_t\pi(\mathbf{x} \oplus \mathbf{e}_\ell(t)) = \lambda_t\pi(\mathbf{x}), \qquad \mathbf{x} \in \mathcal{X},$$

where $\mathbf{x} \oplus \mathbf{e}_\ell(t) = (t(1),\cdots,t(\ell-1),t,t(\ell+1),\ldots,t(n))$ for $\mathbf{x} = (t(1),\cdots,t(n))$. since $\sigma(n)\mu_{t(\ell)}\pi(\mathbf{x}) = \lambda_{t(\ell)}\pi(\mathbf{x}_\ell)$ by the symmetric condition. These are the well-known results originally obtained by [18] and Chandy, Howard and Towsley [8]. We here note that (4.93) fully verifies the insensitivity with respect to the distributions of the amount of work for all types due to 4.10. □

Consider an open or closed queueing network with multi-class Markovian routing whose nodes have symmetric service discipline in the sense of Example 4.18. Then, each node in separation with multi-class Poisson arrivals is quasi-reversible. Hence, from the product form solution for a quasi-reversible network (see, e.g., [9]), if all service requirement distributions are exponential and exogenous arrivals are subject to Poisson processes, then this queueing network has the product form stationary distribution for all type configurations over the network, and satisfies the local balance at each node for each type of customers.

This concludes that the stationary distribution is insensitive with respect to the distributions of the amount of work for all types of customers at each node. This result is usually verified by approximating such distributions by phase types of distributions or by assuming the densities of those distributions. We can again fully verify it by 4.10.

For those product form queueing networks, we can also consider the case that there are negative customers or negative signals at each node as in the condition (4.17d). Similar to the single node case, we can show the product form decomposability by Theorem 4.11. Of course, the macro state distribution can not be insensitive in this case.

4.18 Further insensitivity structure in RGSMP

The product form decomposable RGSMP has insensitive structure not only for the stationary distribution but also for other characteristics. A most prominent feature among them is the conditional mean actual lifetime of a clock given its nominal lifetime, where the nominal lifetime is meant the total amount of lifetime when the clock always advances with unit speed. In RGSMP, speeds of clocks may change, so the actual lifetimes are different from their nominal lifetimes in general. The actual lifetimes are interesting for us since they correspond with the sojourn times of customers in symmetric queues and their networks. We shall show that the mean total sojourn time of a customer arriving at a product form decomposable queueing network is proportional to his total work for service, and its coefficient can be computed.

We first consider the attained sojourn time of an arbitrary fixed clock of a fixed insensitive type $d \in D_g$ in RGSMS. Such a clock is called *tagged*. For this purpose, besides the initial distribution of the RGSMP $\{\mathbf{Z}(t)\}$ given in (4.91), we will consider a further initial distribution which will be specified below and which can be interpreted as a conditional version of that given in (4.91) under the condition that, at time zero, a new (tagged) clock of type $d \in D_g$ is activated.

Let τ^* denote the (total) nominal lifetime of the tagged clock. For $y \leq \tau^*$, let T_y^* be the length of time required by the tagged clock to process y units of its nominal lifetime and let $\ell^*(t)$ denote the site at which this clock is at time $t \leq T_y^*$. Then T_y^* is given by

$$T_y^* = \sup\left\{t > 0 : \int_0^t c_{\mathbf{X}(u)\ell^*(u)} du < y\right\}. \tag{4.94}$$

Throughout this section we assume that the point process N generated by macro state transition instants has finite intensity λ.

We need further notation for describing various point processes arising in connection with stationary RGSMP. Let $N_{(d)}$ be the point process generated by all jump instants at which a new clock of type d is activated. Let $\lambda_{(d)}$ and $P_{(d)}$ denote the intensity of $N_{(d)}$ and the Palm distribution of P with respect to $N_{(d)}$, respectively. Note that $P_{(d)}$ can be interpreted as the conditional probability measure of P given that a clock of type d starts at time 0. For each site $s \in J$, we also introduce the point process N_s generated by all jump instants at which site s gets a new clock of type d, and denote its intensity and the corresponding Palm distribution by λ_s and P_s, respectively. Furthermore, we use the following notation:

$$J_d = \cup_{\mathbf{x}\in\mathcal{X}}\{s \in J(\mathbf{x}) : \gamma_{\mathbf{x}}(s) = d\}, \qquad \mathcal{X}_s = \{\mathbf{x} \in \mathcal{X} : s \in J(\mathbf{x})\}.$$

Since $N_{(d)} = \sum_{s\in J_d} N_s$, from the definition of Palm distribution, we have (see (4.14))

$$\lambda_{(d)}P_{(d)}(C) = \sum_{s\in J_d} \lambda_s P_s(C), \qquad C \in \mathcal{F}. \tag{4.95}$$

By $A^*(t)$ we denote the amount of the nominal lifetime of the tagged clock processed up to time t, i.e.

$$A^*(t) = \int_0^t c_{\mathbf{X}(u)\ell^*(u)} du$$

for every $t \geq 0$ with $A^*(t) < \tau^*$. For $t \geq T_{\tau^*}^*$, we put $A^*(t) \equiv \tau^*$. Furthermore, by $A_s^*(t)$ we denote the amount of the nominal lifetime that a clock of type d, which has been activated at time zero at site s, has consumed up to time $t \geq 0$. Since $P_s(N_{s'}(0) = 1) = \mathbf{1}_{\{s\}}(s')$ for $s, s' \in J$. Hence, (4.95) yields, for $u, y \geq 0$, $s' \in J_d$ and $C \in \mathcal{F}$,

$$\begin{aligned}\lambda_{(d)}P_{(d)}(A^*(u) < y, l^*(0+) = s', C) &= \sum_{s''\in J_d} \lambda_{s''}P_{s''}(A^*(u) < y, l^*(0+) = s', C)\\ &= \lambda_{s'}P_{s'}(A_{s'}^*(u) < y, C)\end{aligned}$$

Thus, by summing up for all possible s' in the above equation, we get the following result.

Lemma 4.15. For $u, y \geq 0$ and $C \in \mathcal{F}$,

$$\lambda_{(d)}P_{(d)}(A^*(u) < y, C) = \sum_{s' \in J_d} \lambda_{s'} P_{s'}(A^*_{s'}(u) < y, C) \, .$$

We denote the nominal lifetime τ^* of the tagged clock by τ^*_s if the tagged clock is created at site s. For $s \in S$, $\mathbf{x} \in \mathcal{X}_s$ and $u, \mathbf{y}_\ell \geq 0$, define the event $C_{\mathbf{x}s}(u, \mathbf{y}_\ell) \in \mathcal{F}$ by

$$C_{\mathbf{x}s}(u, \mathbf{y}_\ell) \equiv \{X(u) = \mathbf{x}, R_{s'}(u) \leq y_{s'}(s' \in J(\mathbf{x}) \setminus \{s\})\} \, ,$$

where $\mathbf{y}_\ell = \{y_{s'}; s' \in J(\mathbf{x}) \setminus \{s\}\}$. Since the probability $P_{s'}(A^*_{s'}(u) < y, \ell^*(u) = s, C_{\mathbf{x}s}(u, \mathbf{y}_\ell) \,|\, \tau^*_{s'})$ does not depend on $\tau^*_{s'}$ on the set $\{\tau^*_{s'} \geq y\} \in \mathcal{F}$, we can write, for $0 \leq z \leq t^* - y$,

$$\begin{aligned} &P_{s'}(A^*_{s'}(u) < y, \ell^*(u) = s, C_{\mathbf{x}s}(u, \mathbf{y}_\ell) \,|\, \tau^*_{s'} = t^*) \\ &\qquad = P_{s'}(A^*_{s'}(u) < y, \ell^*(u) = s, C_{\mathbf{x}s}(u, \mathbf{y}_\ell) \,|\, \tau^*_{s'} = y + z) \, , \end{aligned} \tag{4.96}$$

where $t^* = \sup\{u : 1 - F_d(u) > 0\}$. Moreover, by 4.15, we have

$$\begin{aligned} &\lambda_{(d)} \int_0^\infty P_{(d)}(A^*(u) < y, \ell^*(u) = s, C_{\mathbf{x}s}(u, \mathbf{y}_\ell) | \, \tau^* = z) F_d(dz) \\ &= \sum_{s' \in J_d} \lambda_{s'} \int_0^\infty P_{s'}(A^*_{s'}(u) < y, \ell^*(u) = s, C_{\mathbf{x}s}(u, \mathbf{y}_\ell) | \, \tau^*_{s'} = z) F_d(dz) \, . \end{aligned} \tag{4.97}$$

We are now in a position to prove the next lemma.

Lemma 4.16. Assume that F_d is purely atomic and has a finite number of atoms, i.e. $F_d(x)$ is a step function with a finite number of jumps. Then, for every $\mathbf{x} \in \mathcal{X}$ and $s \in J(\mathbf{x})$ satisfying $\gamma_{\mathbf{x}}(s) = d$, for $0 \leq y \leq t^*$ and for $\mathbf{y}_\ell \geq 0$, we have

$$\begin{aligned} &\mu_d \pi(\mathbf{x}) y \prod_{s' \in J(\mathbf{x}) \setminus \{s\}} F^{(r)}_{\gamma_{\mathbf{x}}(s')}(y_{s'}) \\ &\qquad = \lambda_{(d)} \int_0^\infty P_{(d)}(A^*(u) < y, \ell^*(u) = s, C_{\mathbf{x}s}(u, \mathbf{y}_\ell) \,|\, \tau^* = t^*) du \end{aligned} \tag{4.98}$$

Proof. Let $k_s(t)$ denote the site at which the clock being at time t at site s was originally activated. For $u \leq v$ let $A_s(u, v)$, $\ell_s(u, v)$ and $\tau_s(u, v)$ be the attained sojourn time, the site and the nominal lifetime, respectively, of a clock of type d at time v which started at site s at time u. Let $x \geq 0$, $s, s' \in J_d$, $\mathbf{x} \in \mathcal{X}_s$, and $\mathbf{y}_\ell \geq 0$ be arbitrary but fixed. Then, by the definition of $\ell_{s'}(u, v)$ and $N_{s'}$, we have

$$\begin{aligned} &P(R_s(0) > y, k_s(0) = s', C_{\mathbf{x}s}(0, \mathbf{y}_\ell)) \\ &\qquad = E\left(\int_{-\infty}^0 \mathbf{1}_{\{R_s(0) > y, k_s(0) = s', C_{\mathbf{x}s}(0, \mathbf{y}_\ell), \ell_{s'}(u, 0) = s\}} N_{s'}(du) \right) \\ &\qquad = E\left(\int_{-\infty}^0 \mathbf{1}_{\{R_s(0) > y, C_{\mathbf{x}s}(0, \mathbf{y}_\ell), \ell_{s'}(u, 0) = s\}} N_{s'}(du) \right) . \end{aligned}$$

Moreover, note that, for $u < 0$, $A_{s'}(u,0) + R_s(0) = \tau_{s'}(u,0)$ on the set $\{\ell_{s'}(u,0) = s\}$, and that $A_{s'}(u,0) = A_{s'}(0,-u) \circ \theta_u$, $\tau_{s'}(u,0) = \tau_{s'}(0,-u) \circ \theta_u$ and $\ell_{s'}(u,0) = \ell_{s'}(0,-u) \circ \theta_u$. Then, the last term of the above formula becomes

$$\begin{aligned} &E\left(\int_{-\infty}^{0} \mathbf{1}_{\{A_{s'}(u,0)<\tau_{s'}(u,0)-y, C_{\mathbf{x}s}(0,\mathbf{y}_\ell), \ell_{s'}(u,0)=s\}} N_{s'}(du)\right) \\ &\qquad = E\left(\int_{-\infty}^{0} \mathbf{1}_{\{A_{s'}(0,-u)<\tau_{s'}(0,-u)-y, C_{\mathbf{x}s}(-u,\mathbf{y}_\ell), \ell_{s'}(0,-u)=s\}} \circ \theta_u N_{s'}(du)\right), \end{aligned}$$

which, by 4.6, equals

$$\begin{aligned} &\lambda_{s'} E_{s'}\left(\int_{-\infty}^{0} \mathbf{1}_{\{A_{s'}(0,-u)<\tau_{s'}(0,-u)-y, C_{\mathbf{x}s}(-u,\mathbf{y}_\ell), \ell_{s'}(0,-u)=s\}} du\right) \\ &\qquad = \lambda_{s'} E_{s'}\left(\int_{0}^{\infty} \mathbf{1}_{\{A_{s'}(0,u)<\tau_{s'}(0,u)-y, C_{\mathbf{x}s}(u,\mathbf{y}_\ell), \ell_{s'}(0,u)=s\}} du\right) \\ &\qquad = \lambda_{s'} \int_{0}^{\infty} P_{s'}\left(A_{s'}(0,u) < \tau_{s'}(0,u) - y, C_{\mathbf{x}s}(u,\mathbf{y}_\ell), \ell_{s'}(0,u) = s\right) du, \end{aligned}$$

where $E_{s'}$ denotes the expectation taken with respect to the Palm distribution $P_{s'}$. Thus, from the fact that

$$A_{s'}(0,u) = A^*_{s'}(u), \qquad \tau_{s'}(0,u) = \tau^*_{s'}, \qquad \ell_{s'}(0,u) = s = \ell^*(u) \qquad P_{s'}\text{-a.s.},$$

we get

$$\begin{aligned} &P(R_s(0) > y, k_s(0) = s', C_{\mathbf{x}s}(0,\mathbf{y}_\ell)) \\ &= \lambda_{s'} \int_0^\infty \Big(\int_y^{t^*} P_{s'}(A^*_{s'}(u) < z - y, \ell^*(u) = s, C_{\mathbf{x}s}(u,\mathbf{y}_\ell) |\ \tau^*_{s'} = z) F_d(dz)\Big) du \\ &= \lambda_{s'} \int_y^{t^*} \Big(\int_0^\infty P_{s'}(A^*_{s'}(u) < z - y, \ell^*(u) = s, C_{\mathbf{x}s}(u,\mathbf{y}_\ell) |\ \tau^*_{s'} = t^*) du\Big) F_d(dz), \end{aligned} \quad (4.99)$$

where we have used (4.96) in the last equality of (4.99). Define a function H_d by

$$H_d(y,\mathbf{y}_\ell) = \int_0^\infty P_{(d)}(A^*(u) < y, \ell^*(u) = s, C_{\mathbf{x}s}(u,\mathbf{y}_\ell) \,|\, \tau^* = t^*) du \,.$$

Sum up both sides of (4.99) for all $s' \in J_d$, then (4.97) yields

$$P(R_s(0) > y, C_{\mathbf{x}s}(0,\mathbf{y}_\ell)) = \lambda_{(d)} \int_y^{t^*} H_d(z-y,\mathbf{y}_\ell) F_d(dz) \,. \qquad (4.100)$$

On the other hand, from (4.91), the left-hand side of (4.100) becomes

$$\pi(\mathbf{x}) \overline{F}^{(r)}_d(y) \prod_{s' \in J(\mathbf{x}) \setminus \{s\}} F^{(r)}_{\gamma_{\mathbf{x}}(s')}(y_{s'}) = \mu_d \pi(\mathbf{x}) \prod_{s' \in J(\mathbf{x}) \setminus \{s\}} F^{(r)}_{\gamma_{\mathbf{x}}(s')}(y_{s'}) \int_y^{t^*} (z-y) F_d(dz) \quad (4.101)$$

where $\overline{F}_d^{(r)}(y) = 1 - F_d^{(r)}(y)$. Because of our assumption on F_d, there exist a positive integer n, two sets of positive numbers $\{a_i; i = 1, 2, \dots, n\}$ and $\{p_i; i = 1, 2, \dots, n\}$ satisfying

$$F_d(y) = \sum_{i=1}^{n} p_i \mathbf{1}_{[a_i,\infty)}(y) .$$

Here, we can assume that a_i is increasing in i. Then, from (4.100), (4.101), we get, for $0 \le y \le t^*$,

$$\lambda_{(d)} \sum_{i=1}^{n} p_i H_d((a_i - y)^+, \mathbf{y}_\ell) = \mu_d \pi(\mathbf{x}) \prod_{s' \in J(\mathbf{x}) \setminus \{s\}} F^{(r)}_{\gamma_{\mathbf{x}}(s')}(y_{s'}) \sum_{i=1}^{n} p_i (a_i - y)^+ \quad (4.102)$$

where $y^+ = \max(y, 0)$. Finally, (4.102) implies that, for $0 \le y \le t^*$,

$$\lambda_{(d)} H_d(y, \mathbf{y}_\ell) = \mu_d \pi(\mathbf{x}) y \prod_{s' \in J(\mathbf{x}) \setminus \{s\}} F^{(r)}_{\gamma_{\mathbf{x}}(s')}(y_{s'}) . \quad (4.103)$$

This can be proved in the following way. Consider (4.102) for each sub-interval $(a_{i-1}, a_i]$, where $a_0 = 0$. First, from (4.102) for $y \in (a_{n-1}, a_n]$, we have (4.103) for $0 \le y \le a_n - a_{n-1}$. Then, from (4.102) for $y \in (a_{n-2}, a_{n-1}]$, we have (4.103) for $a_n - a_{n-1} \le y \le \min[a_n - a_{n-2}, 2(a_n - a_{n-1})]$. If $2(a_n - a_{n-1}) < a_n - a_{n-2}$, then, by using the equation just proved, we get (4.103) for $2(a_n - a_{n-1}) \le y \le \min[a_n - a_{n-2}, 3(a_n - a_{n-1})]$. We repeat the argument and eventually get (4.103) for $a_n - a_{n-1} \le y \le a_n - a_{n-2}$. In a similar way we inductively get (4.103) for all the sub-intervals. (4.103) is nothing but (4.98), and therefore the lemma is proved. □

Note that, by (4.94), T^*_y is defined for $0 \le y \le \tau^*$. Now, we extend T^*_y to the whole non-negative half-line by changing the nominal lifetime of the tagged clock to infinity, and denote T^*_y in this case by T^∞_y. Clearly $T^*_y = T^\infty_y$ for $0 \le y \le \tau^*$. The nondecreasing process $\{T^\infty_y; y \ge 0\}$ is called a *attained sojourn time process*.

Analogously, by $\ell^\infty(t)$ we denote the site at which the tagged clock is at time t when its nominal lifetime is changed to infinity. Under the assumption of 4.16, we consider a time change of the RGSMP $\{\mathbf{Z}(t)\}$ by $\{T^\infty_t\}$.

Definition 4.14. Let $\{T^\infty_t; t \ge 0\}$ be the attained sojourn time process for a fixed index $d \in D$. Let $\tilde{X}(t) = X(T^\infty_t)$, $\tilde{\ell}(t) = \ell^\infty(T^\infty_t)$ and $\tilde{R}_s(t) = R_s(T^\infty_t)$. We define a time-changed process $\{\tilde{\mathbf{Z}}(t); 0 \le t < \infty\}$ as

$$\tilde{\mathbf{Z}}(t) = (\tilde{X}(t), \tilde{\ell}(t), \tilde{R}_{s'}(t); s' \in J(\tilde{X}(t)) \setminus \{\tilde{\ell}(t)\}) .$$

This process is said to be a time changed RGSMP concerning the attained lifetime.

Note that $\{\tilde{\mathbf{Z}}(t)\}$ is a Markov process because we can trace its history by using analogous dynamics as for the strong Markov process $\{\mathbf{Z}(t)\}$ and by using the supplementary information $\tilde{\ell}(t)$, which indicates the site at which the tagged clock is at present. For $s \in J_d$, $\mathbf{x} \in \mathcal{X}_s$ and $t, \mathbf{y}_\ell \ge 0$, define the event $\tilde{C}_{\mathbf{x}s}(u, \mathbf{y}_\ell) \in \mathcal{F}$ by

$$\tilde{C}_{\mathbf{x}s}(t, \mathbf{y}_\ell) \equiv \{\tilde{X}(t) = \mathbf{x}, \tilde{R}_{s'}(t) \le y_{s'} (s' \in J(\mathbf{x}) \setminus \{s\})\} .$$

Note that, if we put $v = A^*(u)$ on $\{l^*(u) = s\} \cap C_{\mathbf{x}s}(u, \mathbf{y}_\ell)$, then $dv = c_{\mathbf{x}s} du$ and $T^*_{A^*(u)} = u$ for $c_{\mathbf{x}s} > 0$ while $dv = 0$ for $c_{\mathbf{x}s} = 0$. Hence, by Fubini's theorem and by changing variables from u to $v = A^*(u)$, we have, for $0 \le y \le t^*$,

$$\begin{aligned}
& c_{\mathbf{x}s} \int_0^\infty P_{(d)}(A^*(u) < y, \ell^*(u) = s, C_{\mathbf{x}s}(u, \mathbf{y}_\ell) \,|\, \tau^* = t^*) du \\
&= E_{(d)} \left(\int_0^\infty \mathbf{1}_{\{A^*(u) < y, \ell^*(u) = s, C_{\mathbf{x}s}(u, \mathbf{y}_\ell)\}} c_{\mathbf{x}s} du \middle| \tau^* = t^* \right) \\
&= E_{(d)} \left(\int_0^\infty \mathbf{1}_{\{v < y, l^*(T^*_v) = s, C_{\mathbf{x}s}(T^*_v, \mathbf{y}_\ell)\}} dv \middle| \tau^* = t^* \right) \\
&= E_{(d)} \left(\int_0^y \mathbf{1}_{\{\ell^\infty(T^\infty_v) = s, C_{\mathbf{x}s}(T^\infty_v, \mathbf{y}_\ell)\}} dv \right) \\
&= E_{(d)} \left(\int_0^y \mathbf{1}_{\{\tilde{\ell}(v) = s, \tilde{C}_{\mathbf{x}s}(v, \mathbf{y}_\ell)\}} dv \right) = \int_0^y P_{(d)} \left(\tilde{\ell}(v) = s, \tilde{C}_{\mathbf{x}s}(v, \mathbf{y}_\ell) \right) dv
\end{aligned} \tag{4.104}$$

where the expectation $E_{(d)}$ is taken with respect to $P_{(d)}$. Multiplying both sides of (4.98) by $c_{\mathbf{x}s} \lambda_{(d)}^{-1}$, substituting (4.104) into its right-hand side and differentiating it with respect to y, we get, for $0 \le y \le t^*$,

$$\frac{c_{\mathbf{x}s} \mu_d \pi(\mathbf{x})}{\lambda_{(d)}} \prod_{s' \in J(\mathbf{x}) \setminus \{s\}} F^{(r)}_{\gamma_{\mathbf{x}}(s')}(y_{s'}) = P_{(d)}(\tilde{\ell}(x) = s, \tilde{C}_{\mathbf{x}s}(y, \mathbf{y}_\ell)) \,. \tag{4.105}$$

Hence, $\{\tilde{\mathbf{Z}}(t); 0 \le t < \infty\}$ is a stationary process. By summing up both sides of (4.105) for all possible $s, \mathbf{x}$, we get

$$\lambda_{(d)} = \mu_d \sum_{\mathbf{x} \in \mathcal{X}} \sum_{s \in J(\mathbf{x}) \cap J_d} c_{\mathbf{x}s} \pi(\mathbf{x}) \,.$$

Hence, the left-hand side of (4.105) can be expressed by

$$\pi^*(\mathbf{x}, s) \prod_{s' \in J(\mathbf{x}) \setminus \{s\}} F^{(r)}_{\gamma_{\mathbf{x}}(s')}(y_{s'}). \tag{4.106}$$

where $\pi^*(\mathbf{x}, s)$ is the probability distribution on $\{(\mathbf{x}, s); s \in J_d, \mathbf{x} \in \mathcal{X}_s\}$ defined as

$$\pi^*(\mathbf{x}, s) = \frac{c_{\mathbf{x}s} \pi(\mathbf{x})}{\sum\limits_{\mathbf{x}' \in \mathcal{X}} \pi(\mathbf{x}') \sum\limits_{s' \in J(\mathbf{x}') \cap J_d} c(s', \mathbf{x}')}. \tag{4.107}$$

Thus, we arrive at the following result.

Lemma 4.17. Under the assumption of 4.16, $\{\tilde{\mathbf{Z}}(t); t \ge 0\}$ is a stationary Markov process provided that the initial distribution of $\{\tilde{\mathbf{Z}}(t)\}$ is given by (4.106).

We now remove the assumption of 4.16. For this purpose, we will use a certain continuity property of Markov processes. Let F_d be a general lifetime distribution, and let $\{F_{d,n}\}$ be a sequence of distributions which satisfy the condition of 4.16 and which weakly converge to F_d. Let $\{\mathbf{Z}_n(t);t \geq 0\}$ and $\{\tilde{\mathbf{Z}}_n(t);t \geq 0\}$ be the processes corresponding to $\{\mathbf{Z}(t);t \geq 0\}$ and $\{\tilde{\mathbf{Z}}(t);t \geq 0\}$, respectively, for the RGSMP's with $F_{d,n}$ instead of F_d. We define the initial distribution of $\{\tilde{\mathbf{Z}}_n(t)\}$ by (4.106), in which F_d is replaced by $F_{d,n}$. By 4.17, $\{\tilde{\mathbf{Z}}_n(t)\}$ are stationary Markov processes. $\{\tilde{\mathbf{Z}}_n(t)\}$ and $\{\tilde{\mathbf{Z}}(t)\}$ are self-clocking jump processes as introduced in [26]. We apply Theorem 5.2 of [26] to those processes. Condition (i) of this theorem is clearly satisfied because of (4.90). The stationary one-dimensional distribution of $\{\tilde{\mathbf{Z}}_n(t)\}$ weakly converges to the left-hand side of (4.106), and the transition function at the jump instants of $\{\tilde{\mathbf{Z}}_n(t)\}$ satisfy conditions (ii) and (iii) of Theorem 5.2 of [26], which can easily be verified because we only change the lifetime distributions $F_{d,n}$ (see also Remark 5.2 of [26]). Thus, we get

Theorem 4.12. Assume that RGSMP is product form decomposable, and let $d \in D_g$ be fixed. Then, for a general lifetime distribution F_d, $\{\tilde{\mathbf{Z}}(t);t \geq 0\}$ is a stationary Markov process provided that the initial distribution of $\{\tilde{\mathbf{Z}}(t);t \geq 0\}$ is given by (4.106). Furthermore, combining (4.104) with (4.105) and (4.106), we also have (4.98), namely,

$$\frac{c_{\mathbf{x}s}\pi(\mathbf{x})y}{\sum\limits_{\mathbf{x}'\in\mathcal{X}}\pi(\mathbf{x}')\sum\limits_{s'\in J(\mathbf{x}')\cap J_d} c(s',\mathbf{x}')}\prod_{s'\in J(\mathbf{x})\setminus\{s\}} F^{(r)}_{\gamma_{\mathbf{x}}(s')}(y_{s'})$$
$$= \int_0^\infty P_{(d)}(A^*(u) < y, \ell^*(u) = s, C_{\mathbf{x}s}(u,\mathbf{y}_\ell) \mid \tau^* = t^*)du. \quad (4.108)$$

We have the following verbal interpretation of Theorem 4.12. Under stationarity conditions, given we freeze a randomly chosen type-d clock once it has been started (i.e. putting its nominal lifetime equal to infinity), we observe a stationary process if we look at the remaining system at those times T_y^∞ when the frozen clock has consumed y units of resource, i.e. reached age $y \geq 0$. In particular, the distribution we see when the tagged clock has reached age y is the same for all y and hence, if we draw the age to be reached, blindly from some distribution, e.g. from F_d, we have the same distribution of the process at the time this (random) age is reached. Thus, the next corollary is a direct consequence of Theorem 4.12, i.e. the stationarity of $\{\tilde{\mathbf{Z}}(t)\}$.

Lemma 4.18. Under the conditions of Theorem 4.12, for $\mathbf{x} \in \mathcal{X}$, $s \in J_d$ and $y > 0$, we have

$$P_{(d)}(X(0) = \mathbf{x}, l(0) = s) = P_{(d)}(X(T^*_{\tau^*}-) = \mathbf{x}, l(T^*_{\tau^*}-) = s | \tau^* = y). \quad (4.109)$$

Note that formula (4.109) can be somewhat sharpened: In steady state, at the instants right after the starting of a randomly chosen type-d clock and right before expiring of *that same clock*, the joint distributions of the state $\mathbf{x}$, the site $s \in J(\mathbf{x}) \cap J_d$ on which that clock is found, and the residual lifetimes of the other clocks are both the same.

Theorem 4.12 also yields the following corollary because $\{T_y^\infty; x \geq 0\}$ is completely determined by $\{\tilde{\mathbf{Z}}(t); t \geq 0\}$ (see also Theorem 1 of [13]).

Corollary 4.11. Under the conditions of Theorem 4.12, the attained sojourn time process $\{T_y^\infty; y \geq 0\}$ has stationary increments.

Let $T_y^\infty(\mathbf{x}, s)$ denote the total sojourn time of the system in state $\mathbf{x} \in \mathcal{X}$ while the tagged clock is at site $s \in J(\mathbf{x}) \cap J_d$, until the tagged clock has processed y units of its nominal lifetime, where the nominal lifetime of the tagged clock is assumed to be infinity. Then we have the following result.

Theorem 4.13. Under the conditions of Theorem 4.12, we get, for $y \geq 0$, and for $\mathbf{x} \in \mathcal{X}$ and $s \in J(\mathbf{x}) \cap J_d$,

$$E_{(d)}(T_y^\infty(\mathbf{x}, s)) = \frac{\pi(\mathbf{x})}{\sum_{\mathbf{x}' \in \mathcal{X}} \pi(\mathbf{x}') \sum_{s' \in J(\mathbf{x}') \cap J_d} \mathbf{c}_{\mathbf{x}'\mathbf{s}'}} y, \tag{4.110}$$

and, in particular, by summing up over all possible $\mathbf{x}$ and s,

$$E_{(d)}(T_y^\infty) = \frac{\sum_{\mathbf{x} \in \mathcal{X}} \pi(\mathbf{x}) |J(\mathbf{x}) \cap J_d|}{\sum_{\mathbf{x} \in \mathcal{X}} \pi(\mathbf{x}) \sum_{s \in J(\mathbf{x}) \cap J_d} c_{\mathbf{x}s}} y \tag{4.111}$$

where $|J(\mathbf{x}) \cap J_d|$ denotes the number of elements of the set $J(\mathbf{x}) \cap J_d$.

Proof. Since, for $0 \leq y \leq t^*$,

$$T_y^\infty(\mathbf{x}, s) = T_y^*(\mathbf{x}, s) = \int_0^\infty \mathbf{1}_{\{A^*(u) < x, \ell^*(u) = s, X^*(u) = g\}} du ,$$

(4.108) of Theorem 4.12 yields (4.110) and (4.111). □

Remark 4.17. Note that the right-hand side of (4.110) does not depend on $s \in J(\mathbf{x}) \cap J_d$. Furthermore, if we sum (4.110) up over all $\mathbf{x}, s$ such that $c_{\mathbf{x}s} = 0$, we get a formula for the expected total time during which the tagged clock is interrupted (i.e. stands still) up to the time age y is reached.

In some systems, e.g., in the processor-sharing queue, many clocks of a given type d may run at the same time. Consider the time-changed process with respect to a clock of type d whose lifetime is infinite. Suppose that type-d clocks never run at zero speeds. Then the time-changed process is the RGSMP with macrostates $(\mathbf{x}, s)$, $s \in J_d$, $\mathbf{c} \in \mathcal{X}_s$, and where, in state $(\mathbf{x}, s)$, a clock $s' \in J(\mathbf{x}) \setminus \{s\}$ is running at the speed $c_{\mathbf{x}s'}/c_{\mathbf{x}s}$. Because (4.106) is the stationary distribution of this RGSMP, we can, for instance, consider the successive instants at which another type-d clock gets started, and study the corresponding time-changed process, all from the starting point of the first time-changed process. This new time-changed process lives on the states $[s', (\mathbf{x}, s)]$ with $s \in J(\mathbf{x}) \cap J_d$, $s' \in J(\mathbf{x}) \cap J_d$, $s' \neq s$. Let (4.107) be written as $\pi^*(\mathbf{x}, s) = \frac{1}{k^*} c_{\mathbf{x}s} \pi(\mathbf{x})$. Then the corresponding distribution for the second time-changed process is given by

$$\pi^{**}(s', (\mathbf{x}, s)) = \frac{1}{k^{**}} \frac{c_{\mathbf{x}s'}}{c_{\mathbf{x}s}} \pi^{*}(\mathbf{x}, s)$$

on account of (4.107) applied to the second time-changed process. So

$$\pi^{**}(s', (\mathbf{x},s)) = \frac{c_{\mathbf{x}s'}\pi(\mathbf{x})}{k^{*}k^{**}} = \frac{c_{\mathbf{x}s'}\pi(\mathbf{x})}{\sum\limits_{\mathbf{x}' \in \mathcal{X}} \pi(\mathbf{x}') \sum\limits_{s'' \in J(\mathbf{x}') \cap J_d} c_{\mathbf{x}'s''}}.$$

This would be, in steady state, the probability that, right after the instant of birth of a type-d clock chosen at random while another type-d clock is already running (at site s), the state is $\mathbf{x}$ and that clock is sitting on s'.

Example 4.19 (Symmetric queue). Consider the symmetric queue of Section 4.17. Since there is only one type of customers, $|J(\mathbf{x}) \cap J_d| = n$ for $\mathbf{x} = n$. Hence, by Theorem 4.13, the conditional mean sojourn time of a customer who brings y amount of work is

$$E_d(T_y^{\infty}) = \frac{\sum_{n=1}^{\infty} n\rho^n \prod_{i=1}^{n} \sigma(i)^{-1}}{\sum_{n=1}^{\infty} \rho^n \prod_{i=1}^{n-1} \sigma(i)^{-1}} y,$$

where $\rho = \lambda/\mu$. In particular, $\sigma(n) = a$ for all $n \geq 1$ for some positive constant a, then

$$E_d(T_y^{\infty}) = \frac{\rho}{a(a-\rho)} y.$$

As is expected, the coefficient of the linear function is proportional to the mean queue length. For the product form decomposable network, similar results can be obtained for a given sequence of the amounts of work at visiting nodes of a tagged customer when his route is specified. □

4.19 Bibliographic notes

We briefly discuss about the literature in this chapter. Point processes and Palm measures are now standard in queueing books (see, e.g., [1]). In particular, Baccelli and Bremaud [2] is devoted to this topic and "Palm calculus" was coined there. Historically, the first comprehensive book on this topic for queues was written by Franken, König, Arndt and Schmidt [14]. However, point processes and Palm distributions are old stuff, going back to the ninety-sixties (see, e.g., [30, 21, 22]). There are some other approaches (see, e.g., [6]). The treatments of this topic from Section 4.2 to 4.7 are somehow different from the standard one as in [2]. We more emphasize the symmetric role of time stationary and Palm probability measures. This idea goes back to Miyazawa [23].

The materials in Sections 4.8 and 4.10 are taken from Miyazawa [24, 25]. The rate conservation law and their applications are surveyed in [28]. Example 4.8 is new. Piece-wise deterministic process (PDMP) in Section 4.12 was coined by Davis [10], and detailed in [11]. However, similar types of processes would have been considered long before since they are typical in queueing applications. Our treatments of PDMP is slightly different from those of Davis' as mentioned in Remark 4.5. Generalized semi-Markov process (GSMP) for the insensitivity in the same section has a long history. The earlier literature is Schassberger [31, 32] and Jansen König and Nawrotzki [15]. However, its limitation had been recognized (see, e.g., [3]). Schassberger [33] proposes "relabeling" to relax the limitation. Reallocatable GSMP (RGSMP) was introduced by Miyazawa[27]. It has a similar mechanism to Schassberger's, but allows interruptions.

The stationary equations in Section 4.12 is taken from Miyazawa [26], and those in Section 4.16 from Miyazawa [27]. Symmetric queue and their networks in Section 4.17 is due to Kelly [18, 19]. The locally balanced conditions and product form solutions are largely discussed in the queueing network literature (see, e.g., [4, 8, 9, 34] and references there). Section 4.18 is largely taken from Miyazawa, Schassberger and Schmidt [29], which generalizes the results in [12, 13].

References

1. Asmussen, S. (2003) *Applied Probability and Queues*, Springer, Berlin/ Heidelberg.
2. Baccelli, F., and Bremaud, P. (2003) *Elements of Queueing Theory: Palm Martingale Calculus and Stochastic Recurrences*, Springer, Berlin.
3. Barbour, A.D. (1982) Generalized semi-Markov schemes and open queueing networks. J. Appl. Prob. 19, 469–474.
4. Baskett, A.D., Chandy, K.M., Muntz, R.R., and Palacios, F.G. (1982) Open, closed and mixed networks of queues with different classes of customers. J. ACM 22, 248–260.
5. Billingsley, P. (1995) *Probability and Measure*, Wiley Series in Probability and Statistics, Wiley, New York.
6. Bremaud, P. (1981) *Point Processes and Queues: Martingale Dynamics*, Springer-Verlag, New York.
7. Burke, P.J. (1956) The output of a stationary queueing system, *Operations Research*, 4, 699–704.
8. Chandy,K.M., Howard Jr.J.H. and Towsley,D.F. (1977) Product form and local balance in queueing networks. J. ACM 24, 250–263.
9. Chao, X., Miyazawa, M. and Pinedo, M. (1999) *Queueing Networks: Customers, Signals and Product Form Solutions*, Wiley, Chichester.
10. Davis, M.H.A. (1984) Piecewise-deterministic Markov Process: A general class of Non-diffusion stochastic models, J. R. Statist. Soc. B 46, 353–388.
11. Davis, M.H.A. (1993) *Markov Models and Optimization*, Chapman & Hall, London.
12. Foley, R., and Klutke, G.-A. (1989) Stationary increments in the accumulated work process in processor sharing queues. J. Appl. Prob. 26, 671–677.
13. Foley, R., Klutke, G.-A., and König, D. (1991) Stationary increments of accumulation processes in queues and generalized semi-Markov schemes. J. Appl. Prob. 28, 864–872.
14. Franken, P., König, D., Arndt, U., and Schmidt, V. (1982) *Queues and Point Processes.* J. Wiley & Sons, New York.
15. Jansen, U., König, D., and Nawrotzki, K. (1979) A criterion of insensitivity for a class of queueing systems with random marked point processes. Math. Operationsforsch. Statist., Ser. Optimization 10, 379–403.

16. Kallenberg, O. (2001) *Foundations of Modern Probability*, Second edition, Springer, Yew York.
17. Karatzas, I. and Shreve, S.E. (1998) *Brownian Motion and Stochastic Calculus*, Second edition, Springer, USA.
18. Kelly, F.P. (1976) Networks of queues. Adv. Appl. Prob. 8, 416–432.
19. Kelly, F.P. (1979) *Reversibility and Stochastic Networks.* J. Wiley & Sons, New York.
20. Little, J. (1961) A proof of the queueing formula: $L = \lambda W$, *Operations Research*, 9, 383–387.
21. Matthes, K. (1962) On the theory of queueing processes. Trans. 3rd Prague Conf. Inform. Theory, Statist. Dec. Funct. Random Processes. Prague, 513–528 (in German).
22. Mecke, J. (1967) Stationäre zufülige Masse auf lokalkompakten Abelschen Gruppen, Z. Wahrscheinlich. verw. Geb. 9 , 36–58.
23. Miyazawa, M. (1977) Time and customer processes in queues with stationary inputs, J. Appl. Prob. 14, 349–357.
24. Miyazawa, M. (1983) The derivation of invariance relations in complex queueing systems with stationary inputs. Adv. Appl. Prob. 15, 874–885.
25. Miyazawa, M. (1985) The intensity conservation law for queues with randomly changed service rate, J. Appl. Prob. 22 , 408–418.
26. Miyazawa, M. (1991) The characterization of the stationary distributions of the supplemented self-clocking jump process. Math. Operat. Res. 16, 547–565.
27. Miyazawa, M. (1993) Insensitivity and product-form decomposability of reallocatable GSMP. Adv. Appl. Prob. 25, 415–437.
28. Miyazawa, M. (1994) Rate conservation laws: a survey. Queueing Systems, *15*, 1–58.
29. Miyazawa, M., Schassberger, R. and Schmidt, V. (1995) On the Structure of Insensitive GSMP with Reallocation and with Point-Process Input. Advances in Applied Probability 27, 203–225.
30. Ryll-Nardzewski, C. (1961) Remarks on processes of calls, Proc. of the 4-th Berkeley Symp. on Math. Stat. and Prob. vol. 2, 455–465.
31. Schassberger,R. (1977a) Insensitivity of steady state distributions of generalized semi-Markov processes. PART I. Ann. Prob. 5, 87–99.
32. Schassberger,R. (1977b) Insensitivity of steady state distributions of generalized semi-Markov processes. PART II. Ann. Prob. 6, 85–93.
33. Schassberger, R. (1986) Two remarks on insensitive stochastic models. Adv. Appl. Prob. 18, 791-814.
34. Serfozo, R. (1999) *Introduction to Stochastic Networks*, Springer–Verlag, New York.
35. Wolff, W.R. (1989) *Stochastic Modeling and the Theory of Queues*, Prentice Hall, New York.

Chapter 5
Networks with Customers, Signals, and Product Form Solutions

Xiuli Chao

Abstract In this chapter we present an overview of the latest developments in queueing networks with product form stationary distributions. Under a general framework that allows instantaneous movements, we present sufficient conditions for the network to possess a product form solution. For the case where transitions can involve at most two nodes, we present necessary and sufficient conditions for the network to have a product form solution.

5.1 Introduction

A queueing network is a system consisting of a finite number of *stations* that provide services to *jobs*. The processing stations in the network are typically referred to as *nodes*. Examples of queueing networks include computer systems, manufacturing systems, job shops, airport terminals, railway or highway systems, and telecommunication systems. In these settings, jobs (data, parts or sub-assemblies, customers, planes, vehicles, phone calls, etc.) arrive at the system, and require some form of service (operation executions, assembly processes, machining, airplane take-offs, bridge or toll booth passings, phone conversations, etc.). Queueing network models have been successfully applied in the performance evaluation and optimization of computer systems, communication systems, manufacturing systems and logistic systems. Typical performance measures of practical interest are sojourn time, congestion level, blocking probability, and throughput; and system design and optimization issues include dynamic routing control of jobs (packets), trunk designs, resource allocation, load balancing, or throughput maximization.

Xiuli Chao
Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI 48109
e-mail: xchao@umich.edu

R.J. Boucherie and N.M. van Dijk (eds.), *Queueing Networks: A Fundamental Approach*,
International Series in Operations Research & Management Science 154,
DOI 10.1007/978-1-4419-6472-4_5, © Springer Science+Business Media, LLC 2011

A number of methods have been developed for analyzing queueing networks, each with its own limitations. For example, rapid advances in computers have made it possible to perform large-scale simulations. Still, in addition to the high costs because of the extensive use of computer time and memory, the results of large-scale simulations tend to be application-specific, and are not always useful for detecting general trends in performance measures. Another method for evaluating a queueing system is approximation. However, a theoretical basis is often required to guarantee that an approximation is not far from the real solution. A theoretical analysis of a queueing model is therefore not only important for its own sake; it is also important to complement simulation results and approximations. Such a theoretical analysis involves determining the stationary distribution of the network states, e.g., the number of jobs at each node, from which various performance measures can be derived.

Clearly, a closed form solution for the stationary distribution, if obtainable, is the most preferred. Of the networks with tractable solutions, networks with product form stationary distributions are the ones most researchers have focused on and most applications are based on. Networks with product form solutions have many properties that facilitate their analysis. In this class of networks, in spite of the high level of interaction between the nodes, the joint distribution of all the nodes is the product of the marginal distributions of the individual nodes. Roughly speaking, it implies that the stationary distribution of the network can be obtained by multiplying the stationary distributions of the individual nodes assuming that each node is in isolation and subject to Poisson arrivals. Due to this property, the analysis of a queueing network reduces to the analysis of single node queues, simplifying the applications tremendously. Nevertheless, we shall see that this area is, from a theoretical as well as from a practical point of view, not as narrow as it appears. As a matter of fact, were it not because of Jackson's celebrated product form result and its extensions, applications of queueing networks would most likely not have been as widespread as they are today.

The study of queueing network starts with the celebrated papers of Jackson (1957) (1963). Other work in the nineteen sixties includes Whittle (1968) and Gordon and Newell (1967). These models focus on networks with exponential processing times. Significant breakthrough of queueing network research appeared in mid seventies with the work of Baskett, Chandy, Muntz, and Palacios (1975), and Kelly (1975) (1976), that extend the product form results to networks with arbitrary processing time distributions and multiple classes of jobs. During the nineteen eighties and nineteen nighties researchers extend the theory of queueing networks with batch movements and networks with instantaneous movements and signals, and the representative works in this area are Henderson and Taylor (1990), Gelenbe (1991), Chao and Pinedo (1993), and Chao and Miyazawa (2000). See also the books by Chao, et al. (1999) and Serfozo (1999).

In this chapter we present an overview of the latest developments in queueing networks with tractable solutions. We focus on continuous-time network models. The starting point for the network is a multi-dimensional continue-time Markov chain. This may sound restricted, but it should be noted that, by expanding the state space appropriately, we can approximate a continuous-time network by a continuous-time

Markov chain to any degree of accuracy. In particular, it allows the processing times and interarrival times to have any phase-type distributions (see Neuts 1986). To present the result in a general format we shall start with an abstract framework. We first develop sufficient conditions for the queueing network to possess product form stationary distribution, and for the case with no instantaneous movements we also present the necessary and sufficient conditions for the network to possess product form solution. Numerous examples are given that are covered by the result of the chapter as special cases.

This chapter consists of ten sections. In the following two sections we present the definition of quasi-reversibility for both nodes without triggering and with triggering. In Sections 4 and 5 we introduce networks with quasi-reversible nodes, without and with triggering respectively. Section 6 presents a special class of queueing networks called networks with positive and negative signals as well as their solution. In Section 7 addresses the following question: What is the necessary and sufficient condition for a network to possess a product form stationary distribution, and a complete answer is given to this question for the class of networks that involve simultaneous transitions of at most two nodes. Quasi-reversibility is revisited in Section 8 under the framework of Section 7, and several classes of networks are investigated for which quasi-reversibility is not only a sufficient, but also a necessary condition for product form. Section extends the results to allow customers to randomly change positions at their nodes, both at arrival and customer departure epoches. We conclude with a brief discussion in Section 10.

Sections 2-6 follow Chao and Miyazawa (2000) and Chao, Miyazawa and Pinedo (1999). Sections 7 and 8 follow from Chao, Miyazawa, Serfozo and Takada (1998) and Takada and Miyazawa (1997), see also Chao Chao, Miyazawa and Pinedo (1999). Section 9 extends the model and results in Bonald and Tran (2007).

5.2 Quasi-Reversibility of Queues

Quasi-reversibility is an input-output property of queues. It implies that when the system is in stochastic equilibrium, the future arrival processes, the current state of the system, and the past departure processes are independent.

In conventional queueing models, a job arrives at a system to receive service, and leaves the system after its service is completed. The networks discussed in this chapter include, in addition to conventional jobs, other entities that carry along commands and induce actions at the nodes where they arrive. These entities are, in case they do not trigger instantaneous departures, still referred to as *jobs*. If, however, an arrival has a positive probability of triggering a departure, it is called a *signal*. Thus, the cascading effects of signals may generate throughout the network an arbitrary number of arrivals and departures simultaneously.

It is useful to make a distinction between different classes of jobs and signals. Jobs of different classes may have different characteristics with respect to their processing requirements, their routings through the network, etc. Signals of different classes may carry different messages and may have different effects on the system. When there is no need to make a distinction between them, jobs as well as signals are referred to as *entities*; they may be viewed as different classes of entities.

In the theory of continuous time Markov chain with transition rates $q(x,y), x,y \in \mathcal{S}$, it is conventional to define $q(x,x)$ as $-\sum_{x' \in \mathcal{S}\setminus\{x\}} q(x,x')$. However for the purpose of our study, we forgo this convention and define $q(x,x)$ as a nonnegative number presenting the transition rate from x to itself. This modification allows us to signal such event as arriving entities that cause no change of stage.

In some stochastic systems, such as Example 1 below, both the arrival and the service completion of a regular job result in a transition from $n+1$ to n. Hence, different events may result in the same transition from x to x'. For this reason, we introduce the following notation. Let the system be modeled by a continuous time Markov chain with state space $\mathcal{S}$ and transition rates. $q(x,x'), x,x' \in \mathcal{S}$. For each pair of states (x,x'), we decompose the transition rate function $q(x,x')$ of the queue into three types of rates, namely,

$$\begin{aligned} &q_u^{\mathrm{A}}(x,x'), && u \in T, \\ &q_v^{\mathrm{D}}(x,x'), && v \in T, \\ &q^{\mathrm{I}}(x,x'), && \end{aligned}$$

where T is the set of the classes of arrivals and departures, which is countable. Even though in many queueing systems the classes of arrivals are different from the classes of departures, we use a single index set T because we can take T as the union of both arrival and departure classes. Thus the transition rate of the queue can be written as

$$q(x,x') = \sum_{u \in T} q_u^{\mathrm{A}}(x,x') + \sum_{v \in T} q_v^{\mathrm{D}}(x,x') + q^{\mathrm{I}}(x,x'), \qquad x,x' \in \mathcal{S}. \tag{5.1}$$

These thinned transition rate functions q_u^{A}, q_v^{D} and q^{I} generate the embedded point processes corresponding to class u arrivals, class v departures and the internal transitions, respectively. The first two embedded point processes are often referred to as the *arrival process of class u entities* and the *departure process of class v entities*. The superscripts "A", "D", and "I" stand for "arrival", "departure" and "internal". The internal transition typically represents a change of status of the jobs such as a decrease of their remaining processing times, or, in the case the node contains multiple processing stations, the movements of jobs among the different stations of the node.

If the supports of the rate functions in (5.1) are disjoint, the decomposition above would only have one term. However, we do not make any restriction with regard to their supports. They are distinguished only by the probabilities, i.e., rate decomposition. It should be noted that, even though $q(x,x')$ is said to be decomposed into three types of components (arrival, departure, and internal transition rates) that result in the same transition from x to x', it is the opposite in applications. One is usually given the arrival, departure and internal rates that result in the same transition, and they have to be added up in order to obtain $q(x,x')$.

Example 1. Consider an $M/M/1$ queue with two classes of arrivals. The first class of arrival, denoted by c, represents regular customer, and an arrival of class c increases the number of customer in system by 1. The second class of arrival, denoted by c^- and referred to as negative customer, is a kind of entity whose arrival decreases the number of customer in system, if any, by 1. Thus, $T = \{c, c^-\}$. Service times are exponentially distributed with mean $1/\mu$, and a service completion is classified as class c departure:

$$q_c^{\mathrm{D}}(n,n-1) = \mu, \qquad n = 1,2,\ldots.$$

Regular customers arrive according to a Poisson process with rate α, so

$$q_c^{\mathrm{A}}(n,n+1) = \alpha, \qquad n = 0,1,\ldots.$$

Negative customers arrive according to a Poisson process with rate α^- and reduce the number of customers by 1, we have

$$q_{c^-}^{\mathrm{A}}(n,n-1) = \alpha^-, \qquad n = 1,2,\ldots.$$

Finally, a negative customer that arrives at an empty node simply disappears, thus

$$q_{c^-}^{\mathrm{A}}(0,0) = \alpha^-.$$

Let $q(n,n')$ denote the transition rate of the queue, then its non-zero transition rates are

$$\begin{aligned} q(n,n+1) &= q_c^{\mathrm{A}}(n,n+1), && n \geq 0, \\ q(n,n-1) &= q_c^{\mathrm{D}}(n,n-1) + q_{c^-}^{\mathrm{A}}(n,n-1), && n \geq 1, \\ q(0,0) &= q_{c^-}^{\mathrm{A}}(0,0). \end{aligned}$$

Definition 1. The continuous time Markov chain with transition rate q is called *quasi-reversible with respect to* $\{q_u^{\mathrm{A}}(x,x'); u \in T\}$, $\{q_u^{\mathrm{D}}(x,x'); u \in T\}$ *and* $q^{\mathrm{I}}(x,x')$ if there exist two sets of non-negative numbers $\{\alpha_u; u \in T\}$ and $\{\beta_u; u \in T\}$ such that

$$\sum_{x' \in \mathcal{S}} q_u^{\mathrm{A}}(x,x') = \alpha_u, \qquad x \in \mathcal{S}, u \in T, \tag{5.2}$$

$$\sum_{x' \in \mathcal{S}} \pi(x') q_u^{\mathrm{D}}(x',x) = \beta_u \pi(x), \qquad x \in \mathcal{S}, u \in T, \tag{5.3}$$

where π is the stationary distribution of the Markov chain q.

The non-negative numbers α_u and β_u are often called the arrival rate and departure rate of class u entities.

Quasi-reversibility is a property concerning the arrival and departure processes. In fact, it is often useful to study this property with regard to only a *portion* of the arrival and departure processes. This is particularly true in networks of queues where only some of the departures from a node join another node, while the rest are either absorbed at that node or exit the network. These cases, however, are included in the definition above since one can classify the non-routed arrivals (or departures) as internal transitions.

An alternative definition for quasi-reversibility is the following.

Definition 2. A stationary continuous time Markov chain $\{X(t);t \geq 0\}$ with transition rate q of (5.1) is quasi-reversible if the following two conditions hold.

(i) The $X(t)$ is independent of the arrival process of class u entities subsequent to time t for all $u \in T$.
(ii) The $X(t)$ is independent of the departure process of class u entities prior to time t for $u \in T$.

Before going further, we present an important result, known as Kelly lemma, which will be used numerous times later in this chapter. Its proof can be found, for example, in Kelly (1979).

Lemma 1. (Kelly lemma) For a stationary continuous time Markov chain with state space $\mathcal{S}$ and transition rates $q(x,x')$, if we can find a collection of nonnegative numbers $\tilde{q}(x,x'),x,x' \in \mathcal{S}$ and a collection of positive numbers $\pi(x),x \in \mathcal{S}$, summing to unity, such that

$$\begin{aligned}\sum_{x'\in\mathcal{S}} q(x,x') &= \sum_{x'\in\mathcal{S}} \tilde{q}(x,x'), \qquad x \in \mathcal{S},\\ \pi(x')q(x',x) &= \pi(x)\tilde{q}(x,x'), \qquad x,x' \in \mathcal{S},\end{aligned}$$

then $\tilde{q}(x,x'),x,x' \in \mathcal{S}$ are the transition rates of the reversed process, and $\pi(x), x \in \mathcal{S}$ is the stationary distribution of both processes.

Since different transitions may result in the same change of states, it turns out that a more detailed form of Kelly's lemma is often more convenient to apply. As the relationship between Lemma 1 and the following result is analogous to that of balance equation and detailed balance equation for continuous time Markov chain, we call it detailed Kelly lemma (see Chao, et al. (1999)).

Lemma 2. (Detailed Kelly lemma) Let $q(x,x')$ be the transition rates of a stationary continuous time Markov chain with state space $\mathcal{S}$. Assume that $q(x,x')$ can be decomposed into transition rates $q_\sigma(x,x')$, indexed by $\sigma \in \mathcal{U}$, i.e.,

$$q(x,x') = \sum_{\sigma\in\mathcal{U}} q_\sigma(x,x'), \quad x,x' \in \mathcal{S}.$$

If we can find a collection of nonnegative numbers $\tilde{q}_\sigma(x,x'),x,x' \in \mathcal{S}, \sigma \in \mathcal{U}$, and a collection of positive numbers $\pi(x),x \in \mathcal{S}$, summing to unity, such that

$$\begin{aligned}\sum_{x'\in\mathcal{S}} q_\sigma(x,x') &= \sum_{x'\in\mathcal{S}} \tilde{q}_\sigma(x,x'), \qquad x \in \mathcal{S}, \sigma \in \mathcal{U},\\ \pi(x')q_\sigma(x',x) &= \pi(x)\tilde{q}_\sigma(x,x'), \qquad x,x' \in \mathcal{S},\end{aligned}$$

then

$$\tilde{q}(x,x') = \sum_{\sigma \in \mathcal{U}} \tilde{q}_\sigma(x,x'), \qquad x,x' \in \mathcal{S}$$

are the transition rates of the reversed process, and $\pi(x), x \in \mathcal{S}$ is the stationary distribution of both processes.

Quasi-reversibility is closely related to Poisson flows, as is shown in the following theorem.

Theorem 1. Definitions 1 and 2 are equivalent, and each of them implies that

(a) the arrival process of class $u \in T$ entities are Poisson and the arrival processes of different classes of entities are independent,
(b) the departure process of class $u \in T$ entities are Poisson and the departure processes of different classes of entities are independent.

Proof. The first quasi-reversibility condition (5.2) implies condition (i) of Definition 2 and part (a) of the theorem. On the other hand, condition (i) implies that the left hand side of condition (5.2) is a constant. Denoting this constant by α_u, we obtain (5.2). We next show that condition (5.3) of the first definition implies condition (ii) of the second definition and part (b) of the theorem. To this end, consider the reversed process $X(-t)$. Define $\tilde{q}_u^{\mathrm{A}}$, $\tilde{q}_u^{\mathrm{D}}$ and $\tilde{q}^{\mathrm{I}}$ as

$$\tilde{q}_u^{\mathrm{A}}(x,x') = \frac{\pi(x')}{\pi(x)} q_u^{\mathrm{D}}(x',x),$$

$$\tilde{q}_u^{\mathrm{D}}(x,x') = \frac{\pi(x')}{\pi(x)} q_u^{\mathrm{A}}(x',x),$$

$$\tilde{q}^{\mathrm{I}}(x,x') = \frac{\pi(x')}{\pi(x)} q^{\mathrm{I}}(x',x).$$

Let $\tilde{q}$ be the transition rate function of $X(-t)$. Its transition rates are

$$\tilde{q}(x,x') = \sum_u \tilde{q}_u^{\mathrm{A}}(x,x') + \sum_u \tilde{q}_u^{\mathrm{D}}(x,x') + \tilde{q}^{\mathrm{I}}(x,x').$$

Thus, the time reversed process also represents a queueing model with thinned transitions $\tilde{q}_v^{\mathrm{A}}$, $\tilde{q}_u^{\mathrm{D}}$ and $\tilde{q}^{\mathrm{I}}$. From condition (5.3), we have

$$\sum_{x'} \tilde{q}_u^{\mathrm{A}}(x,x') = \frac{1}{\pi(x)} \sum_{x'} \pi(x') q_u^{\mathrm{D}}(x',x) = \beta_u.$$

This implies that the class u arrival process in $X(-t)$ is Poisson with rate β_u. However, from the definition of $\tilde{q}_u^{\mathrm{A}}$ and the detailed Kelly lemma, a class u arrival in the reversed process $X(-t)$ corresponds to a class u departure in process $X(t)$. Thus we obtain condition (ii) of the second definition. Since a Poisson process reversed in time is Poisson with the same rate, we have (b). Finally, condition (ii) implies that the arrival epochs in the reversed process $X(-t)$ are generated at a constant rate independent of the current state, which gives (5.3).

5.3 Quasi-Reversibility of Queues with Triggered Departures

The quasi-reversibility defined in the last section is concerned with arrival and departure epochs. One prominent feature is that an arrival cannot occur at the same time as a departure. In queueing systems with signals, however, the arrival of a signal may immediately trigger a departure. Thus the definition of quasi-reversibility is not applicable in these cases. In this section we extend the notion of quasi-reversibility to include such simultaneous events.

As before, let q be the transition rate of the node and let it be decomposed into the components $\{q_u^{\mathrm{A}}; u \in T\}$, $\{q_u^{\mathrm{D}}; u \in T\}$ and q^{I} of (5.1). Assume that q admits the stationary distribution π. Furthermore, assume that when a class u entity arrives and induces the state of the node to change from x to x', it instantaneously triggers a class v departure with triggering probability $f_{u,v}(x,x')$, where

$$\sum_{v\in T} f_{u,v}(x,x') \le 1, \qquad u \in T, x, x' \in \mathcal{S}.$$

With probability

$$1 - \sum_{v\in T} f_{u,v}(x,x')$$

the class u arrival does not trigger any departure.

Note that when $\sum_{v\in T} f_{u,v}(x,x') \equiv 0$ for all x and u, the system reduces to that of the previous section with no instantaneous movements.

Definition 3. If there exist two sets of non-negative numbers $\{\alpha_u; u \in T\}$ and $\{\beta_u; u \in T\}$ such that

$$\sum_{x'\in\mathcal{S}} q_u^{\mathrm{A}}(x,x') = \alpha_u, \qquad x \in \mathcal{S}, u \in T, \tag{5.4}$$

$$\sum_{x'\in\mathcal{S}} \pi(x') \left(q_u^{\mathrm{D}}(x',x) + \sum_{v\in T} q_v^{\mathrm{A}}(x',x) f_{v,u}(x',x) \right) = \beta_u \pi(x), \quad x \in \mathcal{S}, u \in T, \tag{5.5}$$

then the queue with signals is said to be quasi-reversible with respect to $\{q_u^{\mathrm{A}}, f_{u,v}; u \in T, v \in T\}$, $\{q_u^{\mathrm{D}}; u \in T\}$, and q^{I}.

As in the last section, quasi-reversibility for queues with signals implies that the arrivals of the different classes of entities form independent Poisson processes, and the departures of different classes of entities, including both triggered and non-triggered departures, also form independent Poisson processes. Moreover, future arrivals and past departures are independent of the current state of the system.

In many applications, triggered and non-triggered departures belong to different classes, i.e.,

$$q_v^{\mathrm{A}}(x,x')f_{v,u}(x,x')q_u^{\mathrm{D}}(x,x') = 0, \qquad \text{for all } x,x' \text{ and } u,v \in T.$$

Let T' and T'' be the sets of the non-triggered and triggered departure classes, respectively, such that

$$T = T' \cup T'', \quad \text{and } T' \cap T'' = \emptyset.$$

Then (5.5) is reduced to

$$\begin{aligned} \sum_{x' \in \mathcal{S}} \pi(x')q_u^{\mathrm{D}}(x',x) &= \beta_u \pi(x), \qquad x \in \mathcal{S}, u \in T', \\ \sum_{x' \in \mathcal{S}} \pi(x') \sum_{v \in T'} q_v^{\mathrm{A}}(x',x)f_{v,u}(x',x) &= \beta_u \pi(x), \qquad x \in \mathcal{S}, u \in T''. \end{aligned}$$

This is equivalent to saying that both the triggered and non-triggered departure processes are independent Poisson with rate β_u for class $u \in T$.

The triggering arrivals and triggered departures are referred to as signals since they pass through a node and change its state instantaneously. The following example illustrates this.

Example 2. Consider an $M/M/1$ queue with two classes of arrivals, denoted by c and s. Class c refers to the regular jobs, and class s refers to signals. When a signal arrives at the node, it triggers a job to depart immediately as a class s departure, provided the queue is not empty upon its arrival. If a signal arrives at an empty queue, nothing occurs and no departure is triggered. The job departures generated by regular processing completions are still classified as class c departures. The decomposed transition rates are

$$\begin{aligned} q_c^{\mathrm{A}}(n,n+1) &= \alpha, \qquad n \geq 0, \\ q_s^{\mathrm{A}}(n,n-1) &= \alpha^-, \qquad n \geq 1, \\ q_s^{\mathrm{A}}(0,0) &= \alpha^-, \\ q_c^{\mathrm{D}}(n,n-1) &= \mu, \qquad n \geq 1. \end{aligned}$$

All other transition rates are zero. By the triggering mechanism, we have

$$\begin{aligned} f_{c,c}(n,n') = f_{c,s}(n,n') &= 0, \qquad n,n' \geq 0, \\ f_{s,s}(n,n-1) &= 1, \qquad n \geq 1. \end{aligned}$$

Since the dynamics of this queue is the same as that of a regular $M/M/1$ queue with arrival rate α and service rate $\mu + \alpha^-$, its stationary distribution π is given by

$$\pi(n) = \left(1 - \frac{\alpha}{\mu + \alpha^-}\right)\left(\frac{\alpha}{\mu + \alpha^-}\right)^n, \qquad n \geq 0.$$

If we set

$$\beta = \frac{\alpha\mu}{\mu + \alpha^-},$$
$$\beta^- = \frac{\alpha\alpha^-}{\mu + \alpha^-},$$

then this system is quasi-reversible with departure rates β and β^-, since,

$$\begin{aligned}
&\sum_{n'} q_c^{\mathrm{A}}(n,n') = \alpha, && n \geq 0,\\
&\sum_{n'} q_s^{\mathrm{A}}(n,n') = \alpha^-, && n \geq 0,\\
&\sum_{n'} \pi(n')\Big(q_c^{\mathrm{D}}(n',n) + \sum_{u=c,s} q_u^{\mathrm{A}}(n',n) f_{u,c}(n',n)\Big) \\
&\qquad = \sum_{n'} \pi(n') q_c^{\mathrm{D}}(n',n) = \beta\pi(n), && n \geq 0,\\
&\sum_{n'} \pi(n')\Big(q_s^{\mathrm{D}}(n',n) + \sum_{u=c,s} q_u^{\mathrm{A}}(n',n) f_{u,s}(n',n)\Big) \\
&\qquad = \sum_{n'} \pi(n') q_s^{\mathrm{A}}(n',n) f_{s,s}(n',n) = \beta^-\pi(n), && n \geq 0.
\end{aligned}$$

This is a very simple system, but many queueing networks with negative signals are generated by this model. □

5.4 Networks of Quasi-Reversible Nodes

In this section we connect N quasi-reversible nodes into a queueing network with Markovian routing mechanisms. The main result is that such a network has a product form solution, i.e., the stationary distribution of the network factorizes into the product of the marginal distributions of the individual nodes.

We consider a queueing network with an arbitrary Markovian routing mechanism and multiple classes of entities. As discussed earlier, entities include both jobs and signals, and their effects on the nodes can be quite general. For instance, the arrival of an entity may decrease the number of jobs or trigger other actions before instantaneously moving to another node. In this section we consider a network structure without signals, i.e., an arrival does not trigger any instantaneous departure. This enables us to give an explicit expression for the network transition rates. The model in this section forms the basis for the network with signals that will be discussed in Section 5. However, when signals are present, the model becomes more involved, and a mathematical expression for the network transition rates becomes complicated without the use of matrix operators.

Suppose the network has N nodes. Each node represents a single processing station, or a cluster of stations (subnetwork). In addition to these nodes we have node 0, which represents the outside world. In this section, the state space $\mathcal{S}_0$ of node 0 is the singleton, i.e., $\mathcal{S}_0 = \{0\}$. Node 0 is a Poisson source, i.e., exogenous entities arrive at the network according to a Poisson process. Even though our main concern is an open network, the arguments can be applied to closed networks as well by simply removing node 0. However, we keep node 0 for consistency. For node $j = 0,1,\ldots,N$, let T_j denote the class of arrival and departure entities at node j. As discussed earlier, we do not make any distinction between arrival and departure classes, even though they may be different. In case they are different we simply let T_j be the union of the arrival and departure classes. For instance, in Example 1, $T_j = \{c, c^-\}$, even though the arrival classes are $\{c, c^-\}$ and the departure class is $\{c\}$.

Let x_j be the state of node j with state space $\mathcal{S}_j$. For node 0, apparently $x_j \equiv 0$. When node j contains a single station, x_j may represent, for instance, the number of each class of jobs as well as their positions in the queue. Since node j may also be a subnetwork, x_j can be more general, e.g., it may represent the number of each class of jobs present at each station as well as their positions at the stations within the node. It may also include the remaining processing times at the node when the processing times are not exponentially distributed. Furthermore, since each node may be a subnetwork, there may be internal transitions within x_j, e.g., job movements between different stations within the same node, or between positions in the same station of node j.

What is the necessary information to construct a queueing network model? A little reflection reveals that we need two types of information: Node (or local) information, i.e., how does each node operate and react to arrivals from other nodes; and inter-node (or global) information, i.e., how are the nodes interconnected.

With regard to node information, we first note that the arrival process at each node of the network is not known before the network is put together. Therefore the arrival transition rate of each node, i.e., q_j^{A}, is not known, nor is it needed for the construction of the queueing network. What we do have to know is what would happen with the node when an arrival occurs. Thus, in order to construct the network, we need for each node the following information.

(i) Arrival effects: The rules according to which the node changes state with the arrival of an entity.
(ii) Departure transition rates: The rate at which the state of the node changes and it may induce the state of another node to change.
(iii) Internal transition rates: The rate at which the state of the node changes and it does not affect the states of other nodes.

For these reasons, we specify each node by a transition *probability* function that describes the changes of state upon arrivals and transition *rate* functions that describe changes of state due to departures and internal transitions. Thus, for node j and an entity of class u, we introduce functions p_{ju}^{A}, q_{ju}^{D} and q_j^{I} on state space $\mathcal{S}_j$.

$p_{ju}^{\mathrm{A}}(x_j, x_j')$ = the probability that a class u arrival at node j changes the state from x_j to x_j', where it is assumed that

$$\sum_{x'_j \in \mathcal{S}_j} p^{\mathrm{A}}_{ju}(x_j, x'_j) = 1, \qquad x_j \in \mathcal{S}_j.$$

$q^{\mathrm{D}}_{ju}(x_j, x'_j) =$ the rate at which class u departures change the state of node j from x_j to x'_j.
$q^{\mathrm{I}}_j(x_j, x'_j) =$ the rate at which internal transitions change the state of node j from x_j to x'_j.

For node 0, we set $p^{\mathrm{A}}_{0,u}(0,0) = 1$, $q^{\mathrm{D}}_{0,u}(0,0) = \beta_{0u}$, and $q^{\mathrm{I}}_j(0,0) = 0$. This implies that exogenous class u entities, i.e., class u departures from node 0, arrive at the network from the outside according to a Poisson process with rate β_{0u}. Note that $p^{\mathrm{A}}_{ju}(x_j, x_j)$ may be positive, i.e., an arrival may not cause a change of state with a positive probability. We refer to p^{A}_{ju} as the *arrival effect function*.

We describe each queue by the three components q^{A}_u, q^{D}_u and q^{I}. If a queue in the network is initially characterized by q^{A}_u, q^{D}_u and q^{I}, then the arrival effect function may be defined as

$$p^{\mathrm{A}}_u(x, x') = \frac{q^{\mathrm{A}}_u(x, x')}{\sum_y q^{\mathrm{A}}_u(x, y)}, \tag{5.6}$$

and q^{D}_u, q^{I} are the departure and internal transition functions. However, unless a node is a separate queue, we assume that it is characterized by p^{A}_u, q^{D}_u and q^{I} because, as discussed earlier, the arrival process at a node of a network depends on the structure of the entire network.

Example 3. Assume that node j of the network is a queue with negative customers, i.e., Example 1. It has two classes of arrivals and a single class of departures, $T_j = \{c, c^-\}$, and is characterized by the following arrival and departure functions:

$$p^{\mathrm{A}}_{jc}(n_j, n'_j) = \begin{cases} 1, \; n'_j = n_j + 1, \\ 0, \text{ otherwise,} \end{cases}$$

$$p^{\mathrm{A}}_{jc^-}(n_j, n'_j) = \begin{cases} 1, \; n'_j = n_j - 1 \geq 0, \\ 0, \text{ otherwise,} \end{cases}$$

$$p^{\mathrm{A}}_{jc^-}(0,0) = 1,$$

$$q^{\mathrm{D}}_{jc}(n_j, n'_j) = \begin{cases} \mu_j, \; n'_j = n_j - 1, n_j \geq 1, \\ 0, \quad \text{otherwise.} \end{cases}$$

There are no class c^- departures and there are no internal transitions, so

$$q^{\mathrm{D}}_{jc^-}(n_j, n'_j) \equiv 0, \qquad n_j, n'_j \geq 0,$$
$$q^{\mathrm{I}}_j(n_j, n'_j) \equiv 0, \qquad n_j, n'_j \geq 0.$$

Node j has a single server with service rate μ_j. When a customer arrives, the number of customers in the node increases by 1, when a negative customer arrives, the number of customer decreases by 1, provided the node is not empty. When a negative customer arrives at an empty node, the state does not change. Note that the arrival process at node j, which is denoted by q^{A}_{ju}, is not given. It is characterized by p^{A}_{ju}, which describes what happens when a class u $(u = c, c^{-})$ entity arrives at the node when it is in state n_j. Also note that p^{A}_{ju} and q^{A}_{ju} in Example 1satisfy (5.6). □

The interactions between the nodes are defined as follows. A class u departure from node j enters node k as a class v arrival with probability $r_{ju,kv}$, and an exogenous class u arrival is routed to node k as a class v arrival with probability $r_{0u,kv}$. It is assumed that

$$\sum_{k=0}^{N} \sum_{v \in T_k} r_{ju,kv} = 1, \qquad j = 0, 1, \ldots, N,\ u \in T_j. \tag{5.7}$$

Note that class u departures from node j leave the network with probability $\sum_{v \in T_0} r_{ju,0v}$. This probability is often denoted by $r_{ju,0}$. In this way, we associate the departures from one node with the arrivals at another.

Let

$$\mathcal{S} = \mathcal{S}_1 \times \mathcal{S}_2 \times \cdots \times \mathcal{S}_N$$

be the product state space. Then,

$$x = (x_1, x_2, \ldots, x_N) \in \mathcal{S}$$

is the state of the network. This network is a continuous time Markov chain with state space $\mathcal{S}$ and transition rate function q, where

$$\begin{aligned} q(x,x') = {} & \sum_{j=0}^{N} \sum_{k=0}^{N} \sum_{u \in T_j} \sum_{v \in T_k} q^{\mathrm{D}}_{ju}(x_j, x'_j)\, r_{ju,kv}\, p^{\mathrm{A}}_{kv}(x_k, x'_k)\, 1[x_\ell = x'_\ell \text{ for all } \ell \neq j,k] \\ & + \sum_{j=0}^{N} q^{\mathrm{I}}_j(x_j, x'_j)\, 1[x_\ell = x'_\ell \text{ for all } \ell \neq j], \end{aligned} \tag{5.8}$$

for $x = (x_1, x_2, \ldots, x_N) \in \mathcal{S}$ and $x' = (x'_1, x'_2, \ldots, x'_N) \in \mathcal{S}$. The first summation on the right hand side of (5.8) represents the state changes due to job transfers from one node to another, and the second summation represents internal state changes. If

$$q^{\mathrm{D}}_{ju}(x_j, x_j) = p^{\mathrm{A}}_{ju}(x_j, x_j) = q^{\mathrm{I}}_j(x_j, x_j) = 0,$$

then the transition rate function (5.8) can be partitioned into disjoint sets:

$$q(x,x') = \begin{cases} \sum_{u \in T} \sum_{v \in T} q^{\mathrm{D}}_{ju}(x_j, x'_j)\, r_{ju,kv}\, p^{\mathrm{A}}_{kv}(x_k, x'_k)\,, & x_\ell = x'_\ell, \quad \text{for all } \ell \neq j,k, \\ q^{\mathrm{I}}_j(x_j, x'_j)\,, & x_\ell = x'_\ell, \quad \text{for all } \ell \neq j, \\ 0\,, & \text{otherwise.} \end{cases}$$

This is a typical situation in a conventional queueing network such as Jackson network. However, it is not true in general. For instance, if a departing job transforms itself into a negative signal and there is no job present at the node where it arrives, then the signal does not have any effect. In this case only the state of the node from which the job departs changes. That is, the transition caused by a signal may result in a change of state that is similar to an internal transition.

We now derive the stationary distribution for the queueing network just constructed. Assuming that each node in isolation is quasi-reversible, we show that the stationary distribution of the network process has product form.

Consider for each node j the following auxiliary process:

$$q_j^{(\alpha_j)}(x_j,x_j') = \sum_{u\in T_j}\left(\alpha_{ju}p_{ju}^{\mathrm{A}}(x_j,x_j') + q_{ju}^{\mathrm{D}}(x_j,x_j')\right) + q_j^{\mathrm{I}}(x_j,x_j'), \quad x_j,x_j' \in \mathcal{S}_j. (5.9)$$

Clearly, $q_j^{(\alpha_j)}(x_j,x_j')$ can be viewed as node j being in isolation, with class $u \in T_j$ entities arriving according to a Poisson process with rate α_{ju}. In general, p_{ju}^{A}, q_{ju}^{D}, and q_j^{I} are allowed to be functions of $\alpha_j = \{\alpha_{ju}; u \in T_j\}$. However, this dependency of α_j is made implicit for simplicity.

Suppose $q_j^{(\alpha_j)}$ has a stationary distribution $\pi_j^{(\alpha_j)}$, i.e.,

$$\pi_j^{(\alpha_j)}(x_j)\left(\sum_{u\in T_j}\left(\alpha_{ju} + \sum_{x_j'\in\mathcal{S}_j} q_{ju}^{\mathrm{D}}(x_j,x_j')\right) + \sum_{x_j'\in\mathcal{S}_j} q_j^{\mathrm{I}}(x_j,x_j')\right)$$
$$= \sum_{u\in T_j}\sum_{x_j'\in\mathcal{S}_j} \pi_j^{(\alpha_j)}(x_j')\left(\alpha_{ju}p_{ju}^{\mathrm{A}}(x_j',x_j) + q_{ju}^{\mathrm{D}}(x_j',x_j)\right) + \sum_{x_j'\in\mathcal{S}_j} \pi_j^{(\alpha_j)}(x_j')q_j^{\mathrm{I}}(x_j',x_j),$$
$$x_j,x_j' \in \mathcal{S}_j. \quad (5.10)$$

This $\pi_j^{(\alpha_j)}$ is expected to be the marginal distribution of node j for some parameters $\alpha_j = (\alpha_{ju}; u \in T_j)$. However, the exact values of $\alpha_1,\ldots,\alpha_N$ are not yet known. Thus, for the time being, the α_j may be regarded as dummy parameters, and their values will be determined later by the traffic equations.

First, note that we always have

$$\sum_{x_j'\in S_j} \alpha_{ju}p_{ju}^{\mathrm{A}}(x_j,x_j') = \alpha_{ju}, \qquad u \in T_j.$$

Hence, quasi-reversibility is equivalent to the property that there exists a set of nonnegative numbers $\{\beta_{ju}; u \in T_j\}$ such that

$$\sum_{x_j'\in\mathcal{S}_j} \pi_j^{(\alpha_j)}(x_j')q_{ju}^{\mathrm{D}}(x_j',x_j) = \beta_{ju}\pi_j^{(\alpha_j)}(x_j), \qquad x_j \in \mathcal{S}_j \quad (5.11)$$

for all $j = 1,2,\ldots,N$ and $v \in T_j$. By (5.11), β_{ju} is determined by

$$\beta_{ju} = \sum_{x_j,x_j'\in\mathcal{S}_j} \pi_j^{(\alpha_j)}(x_j)q_{ju}^{\mathrm{D}}(x_j,x_j'). \quad (5.12)$$

Consider now the network generated by linking nodes $1,2,\ldots,N$ through the routing probability matrix $R=\{r_{ju,kv}\}$, where node j is defined by p^{A}_{ju}, q^{D}_{ju}, and q^{I}_{j}. Assume that class $u\in T_0$ entities arrive at the network from the outside (node 0) at rate β_{0u}, which is given, and that each entity joins node k as a class v entity with probability $r_{0u,kv}$. Let β_{kv} be the average departure rate of class v entities from node k. The average arrival rate of class u entities at node j satisfies

$$\alpha_{ju}=\sum_{k=0}^{N}\sum_{v\in T_k}\beta_{kv}r_{kv,ju},\qquad j=0,1,\ldots,N,\ u\in T_j. \tag{5.13}$$

These equations are referred to as the *traffic equations*. Note that β_{jv} is a non-linear function of α_j, which is determined by (5.12), so the traffic equations are, in general, non-linear in the α_{ju}'s. Finding solutions of (5.13) can be considered a fixed point problem concerning the vector $\alpha=\{\alpha_{ju};j=0,1,\ldots,N,u\in T_j\}$.

Theorem 2. For $(\alpha_0,\alpha_1,\ldots,\alpha_N)$ satisfying (5.11) and (5.13), if each node of the network with transition rate $q_j^{(\alpha_j)}$ is quasi-reversible, then the stationary distribution of the network is

$$\pi(x)=\prod_{j=1}^{N}\pi_j^{(\alpha_j)}(x_j),\qquad x\equiv(x_1,x_2,\ldots,x_N)\in\mathcal{S}. \tag{5.14}$$

Proof. Define distribution π by (5.14). We apply the detailed Kelly lemma to verify that this π is indeed the stationary distribution of the network process q. For convenience, we drop the superscript (α_j). Assume that the time reversed process corresponds to another network with a similar structure. Let

$$\begin{aligned}
\tilde{q}^{\mathrm{D}}_{ju}(x_j',x_j) &= \frac{\pi_j(x_j)\alpha_{ju}p^{\mathrm{A}}_{ju}(x_j,x_j')}{\pi_j(x_j')},\\
\tilde{q}^{\mathrm{I}}_{j}(x_j',x_j) &= \frac{\pi_j(x_j)q^{\mathrm{I}}_{j}(x_j,x_j')}{\pi_j(x_j')},\\
\tilde{p}^{\mathrm{A}}_{ju}(x_j',x_j) &= \frac{\pi_j(x_j)q^{\mathrm{D}}_{ju}(x_j,x_j')}{\pi_j(x_j')\beta_{ju}},\\
\tilde{r}_{ju,kv} &= \frac{\beta_{kv}r_{kv,ju}}{\alpha_{ju}}.
\end{aligned}$$

Note that

$$\sum_{x_j'\in\mathcal{S}_j}\tilde{p}^{\mathrm{A}}_{ju}(x_j,x_j')=1,$$

and

$$\sum_{k=0}^{N}\sum_{v\in T_j}\tilde{r}_{ju,kv}=1.$$

So, we can define a transition rate $\tilde{q}$ for a queueing network process with arrival probability function $\tilde{p}^{\mathrm{A}}_{ju}$, departure rate function $\tilde{q}^{\mathrm{D}}_{ju}$, internal transition rate $\tilde{q}^{\mathrm{I}}_{j}$ and routing probabilities $\tilde{r}_{ju,kv}$.

The detailed Kelly lemma requires the verification of two conditions. To check the first condition, i.e.,

$$\sum_{x'} q(x,x') = \sum_{x'} \tilde{q}(x,x'),$$

note that the state of the network changes only when there is a departure or when there is an internal transition. Thus

$$\begin{aligned}
\sum_{x'} \tilde{q}(x,x') &= \sum_{j=0}^{N} \Big[\sum_{u\in T_j} \sum_{x'_j} \tilde{q}^{\mathrm{D}}_{ju}(x_j,x'_j) + \sum_{x'_j} \tilde{q}^{\mathrm{I}}_{j}(x_j,x'_j) \Big] \\
&= \sum_{j=0}^{N} \left[\sum_{u\in T_j} \sum_{x'_j} \frac{\pi_j(x'_j)\alpha_{ju} p^{\mathrm{A}}_{ju}(x'_j,x_j)}{\pi_j(x_j)} + \sum_{x'_j} \frac{\pi_j(x'_j) q^{\mathrm{I}}_{j}(x'_j,x_j)}{\pi_j(x_j)} \right] \\
&= \sum_{j=0}^{N} \frac{1}{\pi_j(x_j)} \Big[\sum_{u\in T_j} \sum_{x'_j} \pi_j(x'_j) \Big(\alpha_{ju} p^{\mathrm{A}}_{ju}(x'_j,x_j) + q^{\mathrm{D}}_{ju}(x'_j,x_j) \Big) \\
&\qquad + \sum_{x'_j} \pi_j(x'_j) q^{\mathrm{I}}_{j}(x'_j,x_j) - \sum_{u\in T_j} \beta_{ju} \Big] \\
&= \sum_{j=0}^{N} \Big[\sum_{u\in T_j} \sum_{x'_j} q^{\mathrm{D}}_{ju}(x_j,x'_j) + \sum_{x'_j} q^{\mathrm{I}}_{j}(x_j,x'_j) + \sum_{u\in T_j} (\alpha_{ju} - \beta_{ju}) \Big] \\
&= \sum_{j=0}^{N} \Big[\sum_{u\in T_j} \sum_{x'_j} q^{\mathrm{D}}_{ju}(x_j,x'_j) + \sum_{x'_j} q^{\mathrm{I}}_{j}(x_j,x'_j) \Big] \\
&= \sum_{x'} q(x,x'),
\end{aligned}$$

where the third equality follows from (5.11), the fourth equality follows from the fact that π_j is the stationary distribution for q_j, i.e., (5.10), and the last equality follows from the total balance

$$\sum_{j=0}^{N} \sum_{u\in T_j} \alpha_{ju} = \sum_{j=0}^{N} \sum_{u\in T_j} \beta_{ju}, \tag{5.15}$$

which is immediate from the traffic equation.

We next verify the second condition of the detailed Kelly lemma. We decompose $q(x,x')$ and $\tilde{q}(x,x')$ according to the types of transitions. Let $x_j(x'_j)$ be the vector x with its j-th component x_j replaced by x'_j, and denote $(x_j(x'_j))_k(x'_k)$ by $x_{jk}(x'_j,x'_k)$. Note that the types of transitions out of state x under q are $(x,x_j(x'_j))$ and $(x,x_{jk}(x'_j,x'_k))$. The first represents an internal transition at node j and the second a departure from node j, triggering an arrival at node k. In the latter, j may be equal to k, representing a feedback. When this is the case the sequence of transitions is

$$x \xrightarrow{u} x_j(y_j) \xrightarrow{v} x_j(x'_j), \tag{5.16}$$

where u and v are the classes of entities that cause the corresponding transitions. Denote the transition rate for this sequence by

$$q_{ju(y_j),jv}(x,x') = \begin{cases} q^{\mathrm{D}}_{ju}(x_j,y_j) r_{ju,jv} p^{\mathrm{A}}_{jv}(y_j,x_j), & \text{for } x' = x_j(x'_j), \\ 0, & \text{otherwise.} \end{cases}$$

Similarly, for $j \neq k$ and for the sequence of transitions

$$x \xrightarrow{u} x_j(x'_j) \xrightarrow{v} x_{jk}(x'_j,x'_k), \tag{5.17}$$

denote the transition rate by

$$q_{ju,kv}(x,x') = \begin{cases} q^{\mathrm{D}}_{ju}(x_j,x'_j) r_{ju,kv} p^{\mathrm{A}}_{kv}(x_k,x'_k), & \text{for } x' = x_{jk}(x'_j,x'_k), \\ 0, & \text{otherwise.} \end{cases}$$

We similarly define the corresponding rates in the reversed process. Let $\tilde{q}_{ju(y_j),jv}$ denote the rate for the state transitions (5.16), and, for $j \neq k$, let $\tilde{q}_{ju,kv}$ be the rate for the state transitions (5.17). Then, we have

$$q(x,x') = \sum_{j=0}^{N} \left(\sum_{u,v \in T_j} \Big(\sum_{k=0}^{N} q_{ju,kv}(x,x') + \sum_{y_j \in \mathcal{S}_j} q_{ju(y_j),jv}(x,x') \Big) + q^{\mathrm{I}}_j(x,x') \right),$$

$$\tilde{q}(x,x') = \sum_{j=0}^{N} \left(\sum_{u,v \in T_j} \Big(\sum_{k=0}^{N} \tilde{q}_{ju,kv}(x,x') + \sum_{y_j \in \mathcal{S}_j} \tilde{q}_{ju(y_j),jv}(x,x') \Big) + \tilde{q}^{\mathrm{I}}_j(x,x') \right).$$

We now verify the second condition of the detailed Kelly lemma for each sequence of state transitions and each internal transition. The latter is immediate from

$$\begin{aligned} \pi(x_j(x'_j))\tilde{q}^{\mathrm{I}}_j(x'_j,x_j) &= \pi(x_j(x'_j)) \frac{\pi_j(x_j) q^{\mathrm{I}}_j(x_j,x'_j)}{\pi_j(x'_j)} \\ &= \pi(x) q^{\mathrm{I}}_j(x_j,x'_j), \end{aligned}$$

where we have used the fact that $\pi(x)$ is the product of $\pi_j(x_j)$. Similarly,

$$\begin{aligned} \pi(x_j(x'_j))\tilde{q}_{ju(y_j),jv}(x_j(x'_j),x) &= \pi(x_j(x'_j))\tilde{q}^{\mathrm{D}}_{ju}(x'_j,y_j)\tilde{r}_{ju,jv}\tilde{p}^{\mathrm{A}}_{jv}(y_j,x_j) \\ &= \pi(x_j(x'_j)) \frac{\pi_j(y_j)\alpha_{ju} p^{\mathrm{A}}_{ju}(y_j,x'_j)}{\pi_j(x'_j)} \frac{\beta_{jv} r_{jv,ju}}{\alpha_{ju}} \frac{\pi_j(x_j) q^{\mathrm{D}}_{jv}(x_j,y_j)}{\pi_j(y_j)\beta_{jv}} \\ &= \pi(x_j(x'_j)) \frac{\pi_j(x_j)}{\pi_j(x'_j)} q^{\mathrm{D}}_{jv}(x_j,y_j) r_{jv,ju} p^{\mathrm{A}}_{ju}(y_j,x'_j) \\ &= \pi(x) q_{jv,ju(y_j)}(x,x_j(x'_j)). \end{aligned}$$

A similar argument yields, for $j \neq k$,

$$\pi(x_{jk}(x'_j,x'_k))\tilde{q}_{ju,kv}(x_{jk}(x'_j,x'_k),x) = \pi(x)q_{kv,ju}(x,x_{kj}(x'_k,x'_j)).$$

Thus the second condition of Kelly lemma is also satisfied. This completes the proof of the theorem.

In many queueing systems (5.11) holds for a range of α_j. If this is the case the queue is called uniformly quasi-reversible.

Definition 4. Node j, characterized by $\{p^{\mathrm{A}}_{ju}; u \in T_j\}$, $\{q^{\mathrm{A}}_{ju}; u \in T_j\}$ and q^{I}_j, is called *uniformly quasi-reversible* if it is quasi-reversible for all α_j for which the stationary distribution $\pi_j^{(\alpha_j)}$ exists.

If node j is uniformly quasi-reversible, then the departure rate β_{ju} is well defined on the range of α_j for which $\pi_j^{(\alpha_j)}$ exists. Thus it can be considered a function of α_j, and the determination of α_j, for which the marginal distribution $\pi_j^{(\alpha_j)}$ of node j is computed, requires the solution of the non-linear traffic equations. Thus uniform quasi-reversibility is important in computing the stationary distribution of the network. The remaining problem is whether the traffic equations have a solution and how they can be determined. This is a fixed point problem as we stated before on which there exists an extensive literature, thus we will not elaborate it further.

We refer to a network of quasi-reversible nodes as a *quasi-reversible network.* Indeed, as seen from the following corollary, when the entire network is viewed as a system, it is also quasi-reversible.

Corollary 1. In quasi-reversible queueing networks, the class u departure process from node j to the outside is Poisson with rate $\beta_{ju}\sum_{v\in T_0} r_{ju,0v}$. The network is quasi-reversible with respect to arrivals from the outside and departures to the outside.

Proof. From the proof of Theorem 2, the time reversed network process also represents a quasi-reversible network characterized by $\tilde{p}^{\mathrm{A}}_{ju}$, $\tilde{q}^{\mathrm{D}}_{ju}$ and $\tilde{q}^{\mathrm{I}}_j$, and routing probabilities $\tilde{r}_{ju,kv}$. Hence, in the time reversed network, the class u entities arrive at node j from the outside according to a Poisson process with rate

$$\begin{aligned}\sum_{v\in T_0}\alpha_{0v}\tilde{r}_{0v,ju} &= \sum_{v\in T_0}\alpha_{0v}\frac{\beta_{ju}}{\alpha_{0v}}r_{ju,0v}\\ &= \beta_{ju}\sum_{v\in T_0} r_{ju,0v}.\end{aligned}$$

The corollary follows from the fact that the arrivals from the outside in the time reversed process are the departures from the network to the outside in the original network.

However, this does not imply that the departure and arrival processes at each node of the network are Poisson. Actually, it can be shown that the flow on a link is Poisson if and only if it is not part of a cycle. For instance, the flows in between any two nodes in a feedforward network are Poisson.

Example 4. Consider a network with N single-server nodes. Each node has exponentially distributed processing times and two classes of entities: regular customers and negative customers, i.e., the type of node discussed in Example 1. Regular as well as negative customers arrive at node j from the outside according to independent Poisson processes with rates λ_j and λ_j^-. Upon a processing completion at node j, a customer joins node k as a regular customer with probability $r_{jc,kc}$ and as a negative customer with probability r_{jc,kc^-}, $k = 0,1,\ldots,N$. The arrival of a negative customer removes a customer from the node. The state of the network is represented by a vector $n = (n_1,\ldots,n_N)$, where n_j is the number of regular customers at node j, $n_j \in \mathcal{S}_j = \{0,1,\ldots\}$.

Node j, characterized by (5.9), is subject to Poisson arrivals of regular and negative customers with rates α_j and α_j^-. From Example 1, it follows that the stationary distribution π_j of node j is of a geometric form, and the node is uniformly quasi-reversible, and the π_j exists if and only if

$$\alpha_j < \mu_j + \alpha_j^- \,. \tag{5.18}$$

The departure rate from node j is

$$\beta_j = \frac{\alpha_j \mu_j}{\mu_j + \alpha_j^-} \,.$$

The traffic equations are

$$\begin{aligned} \alpha_j &= \lambda_j + \sum_{k=1}^{N} \frac{\alpha_k \mu_k}{\mu_k + \alpha_k^-} r_{kc,jc} \,, \quad j = 1,\ldots,N, \\ \alpha_j^- &= \lambda_j^- + \sum_{k=1}^{N} \frac{\alpha_k \mu_k}{\mu_k + \alpha_k^-} r_{kc,jc^-} \,, \quad j = 1,\ldots,N. \end{aligned}$$

Thus, if the stability condition (5.18) is satisfied, then, by Theorem 2, the stationary distribution of the network, π, is the product of the π_j, i.e.,

$$\pi(n) = \prod_{j=1}^{N} \left(1 - \frac{\alpha_j}{\mu_j + \alpha_j^-}\right) \left(\frac{\alpha_j}{\mu_j + \alpha_j^-}\right)^{n_j} .$$

This is the network first studied by Gelenbe (1991). Gelenbe (1991) introduced the terminology of negative customer in queueing networks, while he called the conventional customers *positive customers*. The reader should not mix this positive customer with a positive signal which we will introduce later. □

5.5 Networks with Signals and Triggered Movements

This section extends the results of the last section to networks with instantaneous movements. Since instantaneous movements are triggered by signals, these networks are often referred to as networks with signals.

Consider a network with N nodes. Each node is a quasi-reversible queue with signals as described in Section 3. Let $\mathcal{S}_j$ be the state space of node j, and let T_j be the set of arrival and departure entity classes, $j = 1,2,\ldots,N$. As discussed in Section 2, we need to specify for each node the transition *probability* functions that describe state changes due to arrivals, and the transition *rate* functions that describe departures and internal state changes. For these, we use the same notation as in last section, i.e., p^{A}_{ju}, q^{D}_{ju} and q^{I}_j. However, since there are instantaneous movements when there is an arrival at a node, we also have to specify the probability functions for the arrivals to induce departures, i.e., the *triggering probability* functions $f_{ju,v}(x_j,x'_j)$. When a class u entity arrives at node j and the state changes from x_j to x'_j, it simultaneously induces a class v departure with triggering probability $f_{ju,v}(x_j,x'_j)$. These probabilities satisfy

$$\sum_{v\in T_j} f_{ju,v}(x_j,x'_j) \le 1, \quad u \in T_j, \quad x_j,x'_j \in \mathcal{S}_j, \quad j = 1,2,\ldots,N.$$

We allow p^{A}_{ju}, $f_{ju,v}$, q^{D}_{ju}, and q^{I}_j to be functions of a nonnegative vector $\alpha_j = \{\alpha_{ju}; u \in T_j\}$, even though this dependency is made implicit for convenience. Also, $p^{\mathrm{A}}_{ju}(x_j,x_j)$ may be positive, i.e., an arrival may, with a positive probability, cause no change of state.

The dynamics of the network is described as follows. Class $u \in T_j$ entities from the outside, i.e., class u departures from node 0, arrive at the network according to a Poisson process with rate β_{0u}, and each is routed to node j as a class v entity with probability $r_{0u,kv}$. A class u departure from node j, either triggered or non-triggered, joins node k as a class v arrival with probability $r_{ju,kv}$, $k = 0,1,\ldots,N$, where

$$\sum_{k=0}^{N}\sum_{v\in T_k} r_{ju,kv} = 1, \quad j = 0,1,\ldots,N,\ u \in T_j.$$

Furthermore, whenever there is a class u arrival at node j, either from the outside or from other nodes, it causes the state of the node to change from x_j to x'_j with probability $p^{\mathrm{A}}_{ju}(x_j,x'_j)$, it also triggers a class v departure with probability $f_{ju,v}(x_j,x'_j)$, and it triggers no departure from node j with probability

$$1 - \sum_{v\in T_j} f_{ju,v}(x_j,x'_j).$$

In this way, we associate the departures, both regular departures and triggered departures, from one node with the arrivals at another.

A distinctive feature of this network is that there are simultaneous arrivals and departures. For instance, if, for nodes $j_1, j_2, \ldots, j_k$ and classes $u_\ell, u'_\ell \in T_\ell$, $\ell = j_1, j_2, \ldots, j_k$,

$$\begin{aligned} p^{\mathrm{A}}_{j_1u_1}(x_{j_1},x'_{j_1}) f_{j_1u_1,u'_1}(x_{j_1},x'_{j_1}) r_{j_1u'_1,j_2u_2} \times \cdots \times p^{\mathrm{A}}_{j_{k-1}u_{k-1}}(x_{j_{k-1}},x'_{j_{k-1}}) \\ \times f_{j_{k-1}u_{k-1},u'_{k-1}}(x_{j_{k-1}},x'_{j_{k-1}}) r_{j_{k-1}u'_{k-1},j_ku_k} p^{\mathrm{A}}_{j_ku_k}(x_{j_k},x'_{j_k}) > 0, \end{aligned}$$

then a class u_1 arrival at node j_1 will simultaneously create arrivals at nodes $j_1, j_2, \ldots, j_k$, and change the states of these nodes to $x'_{j_1}, x'_{j_2}, \ldots, x'_{j_k}$ with a positive probability. Note that the same node may be visited several times on this route, and as a result, the state of the node may change a number of times at one point in time.

Let

$$X(t) = (X_1(t), X_2(t), \ldots, X_N(t))$$

denote the state of the network at time t, with $X_j(t)$ being the state of node j. Then $X(t)$ is a Markov process on the state space $\mathcal{S}$.

The following technical assumption has to be made in order to avoid an infinite number of visits at a node at one time epoch. For any sequence of nodes $j_1, j_2, \ldots, j_\ell$, with arrival classes $u_1, \ldots, u_\ell$, and departure classes $u'_1, \ldots, u'_\ell$, and for any sequence of network states $x, x_1, x_2, \ldots, x_\ell$,

$$\lim_{\ell \to \infty} p_x((j_1 u_1, u'_1, x_1), \ldots, (j_\ell u_\ell, x_\ell)) f_{j_\ell u_\ell, u'_\ell}(x_{\ell-1}, x_\ell) = 0. \tag{5.19}$$

In most applications this assumption is easily verified. For instance, if instantaneous movements always decrease the numbers of jobs at the nodes and stop propagating when they arrive at empty nodes, the network will be empty after a finite number of steps, so the sequence of nodes to be visited is of finite length. On the other hand, if each visit increases the number of jobs, then, without the assumption above, the network will, with positive probability, explode at a single time epoch. Let q denote the transition rate function of the Markov process $X(t)$.

As in the earlier section, we need to first consider each individual node j with an auxiliary transition rate $q_j^{(\alpha_j)}$ to compute the stationary distribution of the network, where

$$q_j^{(\alpha_j)}(x_j, x'_j) = \sum_{u \in T_j} \left(\alpha_{ju} p^{\mathrm{A}}_{ju}(x_j, x'_j) + q^{\mathrm{D}}_{ju}(x_j, x'_j) \right) + q^{\mathrm{I}}_j(x_j, x'_j), \quad x_j, x'_j \in \mathcal{S}_j \tag{5.20}$$

The $\alpha_j = (\alpha_{ju}; u \in T_j)$ are considered dummy parameters and their values are to be determined by the traffic equations. Assume $q_j^{(\alpha_j)}$ has a stationary distribution $\pi_j^{(\alpha_j)}$, $j = 1, 2, \ldots, N$. We now require that the nodes with signals be quasi-reversible. Note that $q^{\mathrm{A}}_{ju} \equiv \alpha_{ju} p^{\mathrm{A}}_{ju}$ satisfies condition (5.5) automatically. So the quasi-reversibility is equivalent to the existence of non-negative numbers $\{\beta_{iu}; u \in T_j\}$ for all $j = 1, 2, \ldots, N$ such that

$$\sum_{x'_j \in \mathcal{S}_j} \pi_j^{(\alpha_j)}(x'_j) \Big(q^{\mathrm{D}}_{ju}(x'_j, x_j) + \sum_{v \in T_j} \alpha_{jv} p^{\mathrm{A}}_{jv}(x'_j, x_j) f_{jv,u}(x'_j, x_j) \Big) = \beta_{ju} \pi_j^{(\alpha_j)}(x_j)$$

$$u \in T_j,\ x_j \in \mathcal{S}_j. \tag{5.21}$$

Since α_{ju} and β_{iu} are the arrival and departure rates of class u entities at node j, the traffic equations

$$\alpha_{ju} = \sum_{k=0}^{N} \sum_{v \in T_k} \beta_{kv} r_{kv,ju}, \qquad j = 0,1,\ldots,N, \quad u \in T_j$$

have to be satisfied. We need the following condition to ensure that the network process is regular:

$$\sum_{j=1}^{N} \sum_{x_j \in \mathcal{S}_j} \pi_j^{(\alpha_j)}(x_j) \sum_{x'_j \in \mathcal{S}_j} q_i^{(\alpha_j)}(x_j)(x_j,x'_j) < \infty. \tag{5.22}$$

A simple sufficient condition for (5.22) is

$$\sum_{u \in T_j} \left(\alpha_{ju} + \beta_{ju} + \sum_{x'_j \in \mathcal{S}_j} q_j^{\mathrm{I}}(x_j)(x_j,x'_j) \right) < \infty, \qquad \text{for all } j = 1,\ldots,N,$$

which is satisfied by all the examples in this chapter.

The following result for networks with signals is an extension of Theorem 2 for networks without instantaneous movements.

Theorem 3. If each node of the network is a quasi-reversible queue with signals, i.e., equation (5.21) is satisfied, and if α_j, $j = 1,\ldots,N$ are the solutions of the traffic equations (5.13), then the queueing network with signals has the product form stationary distribution

$$\pi(x) = \prod_{j=1}^{N} \pi_j^{(\alpha_j)}(x_j), \qquad x \equiv (x_1,x_2,\ldots,x_N) \in \mathcal{S}. \tag{5.23}$$

Proof. We use the detailed Kelly lemma. For convenience we drop the superscript (α_j). Assume that the reversed process corresponds to a similar network that is characterized by

$$\begin{aligned}
\tilde{q}^{\mathrm{D}}_{ju}(x'_j,x_j) &= \frac{\pi_j(x_j)\alpha_{ju} p^{\mathrm{A}}_{ju}(x_j,x'_j)\Big(1 - \sum_v f_{ju,v}(x_j,x'_j)\Big)}{\pi_j(x'_j)} \\
\tilde{q}^{\mathrm{I}}_j(x'_j,x_j) &= \frac{\pi_j(x_j) q^{\mathrm{I}}_j(x_j,x'_j)}{\pi_j(x'_j)}, \\
\tilde{p}^{\mathrm{A}}_{ju}(x'_j,x_j) &= \frac{\pi_j(x_j)\Big(q^{\mathrm{D}}_{ju}(x_j,x'_j) + \sum_v \alpha_{jv} p^{\mathrm{A}}_{ju}(x_j,x'_j) f_{jv,u}(x_j,x'_j)\Big)}{\pi_j(x'_j)\beta_{ju}} \\
\tilde{f}_{jv,u}(x'_j,x_j) &= \frac{\alpha_{jv} p^{\mathrm{A}}_{jv}(x_j,x'_j) f_{jv,u}(x_j,x'_j)}{q^{\mathrm{D}}_{ju}(x_j,x'_j) + \sum_v \alpha_{jv} f_{jv,u}(x_j,x'_j)}, \\
\tilde{r}_{ju,kv} &= \frac{\beta_{kv} r_{kv,ju}}{\alpha_{ju}}.
\end{aligned}$$

Because of the quasi-reversibility condition (5.21) and the traffic equations (5.13), $\tilde{p}^{\mathrm{A}}_{ju}(y_j,x_j)$ and $\tilde{r}_{ju,kv}$ are indeed probabilities. Thus they determine a queueing network with instantaneous movements.

To apply the detailed Kelly lemma, we need to verify two conditions. To check the first condition, i.e.,

$$\sum_{x'} q(x,x') = \sum_{x'} \tilde{q}(x,x'),$$

we first note that changes in the network state are initiated by either a departure or an internal transition, and terminated after a finite number of transitions. The latter is ensured by condition (5.19), which guarantees that the network process is well defined. Thus, similar to the proof of Theorem 2, using the quasi-reversibility condition (5.21), the global balance of each node (5.10) and the total balance (5.15), we obtain

$$\begin{aligned}
\sum_{x'} \tilde{q}(x,x') &= \sum_{j=0}^{N}\Big[\sum_{u\in T_j}\sum_{x'_j}\tilde{q}^{\mathrm{D}}_{ju}(x_j,x'_j)+\sum_{x'_j}\tilde{q}^{\mathrm{I}}_{j}(x_j,x'_j)\Big]\\
&= \sum_{j=0}^{N}\Bigg[\sum_{u\in T_j}\sum_{x'_j}\frac{\pi_j(x'_j)\alpha_{ju}p^{\mathrm{A}}_{ju}(x'_j,x_j)\Big(1-\sum_{v\in T_j} f_{ju,v}(x'_j,x_j)\Big)}{\pi_j(x_j)}\\
&\qquad +\sum_{x'_j}\frac{\pi_j(x'_j)q^{\mathrm{I}}_{j}(x'_j,x_j)}{\pi_j(x_j)}\Bigg]\\
&= \sum_{j=0}^{N}\Big[\Big(\sum_{u\in T_j}\sum_{x'_j}\pi_j(x'_j)\alpha_{ju}p^{\mathrm{A}}_{ju}(x'_j,x_j)+\sum_{u\in T_j}\sum_{x'_j}\pi_j(x'_j)q^{\mathrm{D}}_{ju}(x'_j,x_j)\\
&\qquad +\sum_{x'_j}\pi_j(x'_j)q^{\mathrm{I}}_{j}(x'_j,x_j)\Big)\Big/\pi_j(x_j)-\beta_{ju}\Big]\\
&= \sum_{j=0}^{N}\Big[\sum_{u\in T_j}\sum_{x'_j}q^{\mathrm{D}}_{ju}(x_j,x'_j)+\sum_{x'_j}q^{\mathrm{I}}_{j}(x_j,x'_j)+\alpha_{ju}-\beta_{ju}\Big]\\
&= \sum_{j=0}^{N}\Big[\sum_{u\in T_j}\sum_{x'_j}q^{\mathrm{D}}_{ju}(x_j,x'_j)+q^{\mathrm{I}}_{j}(x_j,x'_j)\Big]\\
&= \sum_{x'} q(x,x'),
\end{aligned}$$

To check the second condition of the detailed Kelly lemma, we need to decompose the transition functions $q(x,x')$ and $\tilde{q}(x,x')$, such that

$$\pi(x)q_{\sigma}(x,x') = \pi(x')\tilde{q}_{\tilde{\sigma}}(x',x) \tag{5.24}$$

is satisfied for each decomposed transition functions $q_{\sigma}(x,x')$ and $\tilde{q}_{\tilde{\sigma}}(x,x')$, where σ is the index for a sequence of simultaneous transitions, and $\tilde{\sigma}$ is the index for the reversed sequence of σ. Clearly, a transition for this network is either an internal transition at a node, or a transition involving at least two nodes (possibly node 0, the outside). A transition involving two or more nodes takes the following form: The network starts out in state x, a class u entity departs from node j ($j = 0, 1, \ldots, N$) and goes to node j_1 as a class u_1 entity, changes the state of the network to x_1, and triggers at the same time a class u'_1 departure from node j_1; it then goes to node j_2 as a class u_2 entity, changes the state of the network to x_2, and triggers a class u'_2 departure, etc. The string transition ends when a class v entity arrives at a node k that does not trigger any instantaneous departure, and the state changes to x'. Let σ denote this sequence. For the reversed process, $\tilde{\sigma}$ represents the reversed sequence of σ which initiates with a class v departure at node k when its state is x', triggering a string of transitions, and ends with a class u arrival at node j that does not trigger any instantaneous departure and the state of the network changes to x.

We now verify the second condition of the detailed Kelly lemma. For an internal transition, (5.24) is easily seen to be satisfied by the definition of $\tilde{q}^{\mathrm{I}}_j$. For transitions that involve more than one node, we consider here only the case of a string transition that involves three nodes.

$$\begin{aligned}
&\pi(x)q_{\sigma}(x,x')\\
&\quad= \pi_j(x_j)\pi_{j_1}(x_{j_1})\pi_k(x_k)q^{\mathrm{D}}_{iu}(x_j,x'_j)r_{ju,j_1u_1}p^{\mathrm{A}}_{j_1u_1}(x_{j_1},x'_{j_1})f_{j_1u_1,u'_1}(x_{j_1},x'_{j_1})\\
&\qquad\qquad\qquad \times r_{j_1u'_1,kv}p^{\mathrm{A}}_{kv}(x_k,x'_k)\Big(1-\sum_{w\in T_k} f_{kv,w}(x_k,x'_k)\Big)\prod_{\ell\neq j,j_1,k}\pi_\ell(x_\ell).
\end{aligned}$$

Similarly the right hand side is

$$\begin{aligned}
&\pi(x')\tilde{q}_{\tilde{\sigma}}(x',x)\\
&\quad= \pi_k(x'_k)\pi_{j_1}(x'_{j_1})\pi_j(x'_j)\tilde{q}^{\mathrm{D}}_{jv}(x'_j,x_j)\tilde{r}_{kv,j_1u'_1}\tilde{p}^{\mathrm{A}}_{j_1u'_1}(x'_{j_1},x_{j_1})\tilde{f}_{j_1u'_1,u_1}(x'_{j_1},x_{j_1})\\
&\qquad\qquad\qquad \times \tilde{r}_{j_1u_1,ju}\tilde{p}^{\mathrm{A}}_{ju}(x'_j,x_j)\Big(1-\sum_{w\in T_j} \tilde{f}_{ju,w}(x'_j,x_j)\Big)\prod_{\ell\neq j,j_1,k}\pi_\ell(x_\ell).
\end{aligned}$$

It is straightforward to verify that these two terms are equal. In case the string contains more than 3 nodes, (5.24) can be verified in a similar way. This completes the proof of Theorem 3.

The following example extends Example 4 by including triggered movements throughout the network. Clearly, if $r_{js,ks} =$ for all j and k, then it reduces to the model of Example 4.

Example 5. Consider a queueing network with jobs and negative signals, as described in Example 2. When a job completes its processing at node j, it goes to node k as a regular job with probability $r_{jc,kc}$, and as a negative signal with probability $r_{jc,ks}$, $k = 0, 1, \ldots, N$. When a negative signal arrives at node j, it induces a job, if there is one present, to depart. The job then joins node k as a regular job with probability $r_{js,kc}$, and as a negative signal with probability $r_{js,ks}$, $k = 0, 1, \ldots, N$. Let λ_j and λ^-_j be the exogenous arrival rates of jobs and signals at node j. The state of the network is given by vector $n = (n_1, \ldots, n_N)$, where n_j is the number of jobs at node j.

Node j, defined by (5.20), is a queue discussed in Example 2, with service rate μ_j, job arrival rate α_j, and signal arrival rate α_j^-. It is quasi-reversible with departure rates of jobs and signals given by

$$\beta_j = \frac{\alpha_j \mu_j}{\mu_j + \alpha_j^-}, \qquad \beta_j^- = \frac{\alpha_j \alpha_j^-}{\mu_j + \alpha_j^-}.$$

Thus the traffic equations are

$$\alpha_j = \lambda_j + \sum_{k=1}^{N} \frac{\alpha_k \mu_k}{\mu_k + \alpha_k^-} r_{kc,jc} + \sum_{k=1}^{N} \frac{\alpha_k \alpha_k^-}{\mu_k + \alpha_k^-} r_{ks,jc}\,, \tag{5.25}$$

$$j = 1, 2, \ldots, N,$$

$$\alpha_j^- = \lambda_j^- + \sum_{k=1}^{N} \frac{\alpha_k \mu_k}{\mu_k + \alpha_k^-} r_{kc,js} + \sum_{k=1}^{N} \frac{\alpha_k \alpha_k^-}{\mu_k + \alpha_k^-} r_{ks,js}\,, \tag{5.26}$$

$$j = 1, 2, \ldots, N.$$

Suppose these traffic equations have positive solutions α_j, α_j^- such that

$$\frac{\alpha_j}{\mu_j + \alpha_j^-} < 1, \qquad j = 1, 2, \ldots, N.$$

Since each node is uniformly quasi-reversible, applying Theorem 3 yields the stationary distribution

$$\pi(n) = \prod_{j=1}^{N} \left(1 - \frac{\alpha_j}{\mu_j + \alpha_j^-}\right) \left(\frac{\alpha_j}{\mu_j + \alpha_j^-}\right)^{n_j}.$$

□

5.6 Networks with Positive and Negative Signals

In this section we apply Theorem 3 to a queueing network with two types of signals: positive signals and negative signals. The first subsection considers the case of a single class of positive signals and a single class of negative signals, and the second subsection considers multiple classes of positive and negative signals. These models include most networks with batch movements as special cases.

5.6.1 Single Class of Positive and Negative Signals

Consider a network with N nodes and a single server at each node. There are three classes of entities: *jobs, positive signals*, and *negative signals*, denoted by $T = \{c, s^+, s^-\}$. Class c refers to the regular jobs; their arrivals do not trigger any instantaneous movements. Positive and negative signals, however, represent signaling mechanisms that induce immediate transitions at the nodes where they arrive. Assume that the state of the network is $n = (n_1, n_2, \ldots, n_N)$, where n_j is the number of jobs at node j, $j = 1, 2, \ldots, N$. The arrival of a positive signal at a node increases the number of jobs at that node by 1, and then leaves immediately for another node. The arrival of a negative signal at a node triggers a job to depart, provided the node is not empty upon its arrival. A negative signal disappears when it arrives at an empty node.

Assume that jobs arrive from the outside at node j according to a Poisson process with rate λ_j, and positive and negative signals arrive from the outside at node j according to Poisson processes with rates λ_j^+ and λ_j^-. Node j, $j = 1, \ldots, N$, has exponential processing times with rate μ_j.

Upon a processing completion at node j, a job leaves for node k as a regular job with probability $r_{jc,kc}$, as a positive signal with probability r_{jc,ks^+}, as a negative signal with probability r_{jc,ks^-}, and it leaves the network with probability $r_{jc,0}$, where

$$\sum_{k=1}^{N} (r_{jc,kc} + r_{jc,ks^+} + r_{jc,ks^-}) + r_{jc,0} = 1, \quad j = 1, \ldots, N.$$

When a positive signal arrives at node j, either from the outside or from another node, it adds one job and then leaves immediately for node k as a regular job with probability $r_{js^+,kc}$, as a positive signal with probability r_{js^+,ks^+}, as a negative signal with probability r_{js^+,ks^-}, and it leaves the network with probability $r_{js^+,0}$, where

$$\sum_{k=1}^{N} (r_{js^+,kc} + r_{js^+,ks^+} + r_{js^+,ks^-}) + r_{js^+,0} = 1, \quad j = 1, \ldots, N.$$

Finally, when a negative signal arrives at node j, either from the outside or from another node, it triggers a job, if any, to depart. The departing job goes to node k as a regular job with probability $r_{js^-,kc}$, as a positive signal with probability r_{js^-,ks^+}, as a negative signal with probability r_{js^-,ks^-}, and it leaves the network with probability $r_{js^-,0}$, where again

$$\sum_{k=1}^{N} (r_{js^-,kc} + r_{js^-,ks^+} + r_{js^-,ks^-}) + r_{js^-,0} = 1, \quad j = 1, \ldots, N.$$

As indicated earlier, a negative signal that arrives at an empty node is assumed to be lost. We refer to this model as *a network with positive and negative signals*.

In this network there can be any number of job additions or deletions at various nodes of the network at the same point in time. For instance, if there is a sequence $j_1, j_2, \ldots, j_k$ such that

$$r_{j_1 s^+, j_2 s^+} r_{j_2 s^+, j_3 s^+} \cdots r_{j_{k-1} s^+, j_k s^+} > 0,$$

then the arrival of a positive signal at node j_1 would, with positive probability, add one job at each one of nodes $j_1, j_2, \ldots, j_k$. There can also be batch arrivals at node j if $r_{js^+,js^+} > 0$, and the arrival of a positive signal can add a batch of random size at a number of nodes. Unlike in networks with negative signals, in which a signal may be interrupted on its route once it hits an empty node, a positive signal in this model will never be interrupted. It disappears only when it is transformed into another class of entity or when it leaves the network.

To ensure that the network is stable, i.e., it will not be overloaded, we have to exclude the case that a positive signal from any node generates an infinite number of jobs in the network at one point in time. Hence we make the following technical assumption in order to ensure that the stochastic process is regular: The Markov chain with state space $\{0, 1, \ldots, N\}$ and transition probabilities $p_{\cdot,\cdot}$ given by

$$\begin{aligned}
&p_{j,k} = r_{js^+,ks^+}, && j,k = 1, \ldots, N,\\
&p_{j,0} = 1 - \sum_{k=1}^{N} r_{js^+,ks^+} - \sum_{k=0}^{N} r_{js^+,kc}, && j = 1, \ldots, N,\\
&p_{0,0} = 1,
\end{aligned}$$

has only one recurrent state 0.

We are interested in the stationary probability of this network. However, it is known that such networks do not have closed form solutions. In the following theorem, we modify the network process so as to obtain a product form solution for the network. This modification may appear artificial, and it is introduced purely to obtain the uniform quasi-reversibility of each node so that Theorem 3 can be applied. However, under some conditions the product form solution serves as a stochastic upper bound for the original network.

Suppose the following traffic equations have a nonnegative solution $\{\alpha_j; j = 1, \ldots, N\}$, $\{\alpha_j^+; j = 1, \ldots, N\}$, and $\{\alpha_j^-; j = 1, \ldots, N\}$:

$$\alpha_j = \lambda_j + \sum_{k=1}^{N} \rho_k \mu_k r_{kc,jc} + \sum_{k=1}^{N} \rho_k \alpha_k^- r_{ks^-,jc} + \sum_{k=1}^{N} \rho_k^{-1} \alpha_k^+ r_{ks^+,jc}, \tag{5.27}$$

$$\alpha_j^+ = \lambda_j^+ + \sum_{k=1}^{N} \rho_k \mu_k r_{kc,js^+} + \sum_{k=1}^{N} \rho_k \alpha_k^- r_{ks^-,js^+} + \sum_{k=1}^{N} \rho_k^{-1} \alpha_k^+ r_{ks^+,js^+}, \tag{5.28}$$

$$\alpha_j^- = \lambda_j^- + \sum_{k=1}^{N} \rho_k \mu_k r_{kc,js^-} + \sum_{k=1}^{N} \rho_k \alpha_k^- r_{ks^-,js^-} + \sum_{k=1}^{N} \rho_k^{-1} \alpha_k^+ r_{ks^+,js^-}, \tag{5.29}$$

for $j = 1, \ldots, N$, where

$$\rho_j = \frac{\alpha_j + \alpha_j^+}{\mu_j + \alpha_j^-}.$$

Now modify the network with positive and negative signals such that whenever node j is empty, a Poisson departure process of positive signals is activated with rate $\rho_j^{-1}\alpha_j^+$.

Theorem 4. If the solution of the traffic equations satisfy $\rho_j < 1$ for $j = 1, \ldots, N$, the modified network described above has the product form stationary distribution

$$\pi(n_1, \ldots, n_N) = \prod_{j=1}^{N} (1 - \rho_j)\rho_j^{n_j}. \tag{5.30}$$

Proof. We use Theorem 3 to prove the result. It suffices to verify the quasi-reversibility of each node when it is in isolation and subject to Poisson arrivals of jobs and signals. Let a processing completion at node j be classified as a class c departure, and departures triggered by positive and negative signals as class s^+ and s^- departures, respectively. Then, node j is characterized by

$$\begin{aligned}
&p_{jc}^{\text{A}}(n_j, n_j + 1) = 1, && n_j \geq 0,\\
&p_{js^+}^{\text{A}}(n_j, n_j + 1) = 1, && n_j \geq 0,\\
&p_{js^-}^{\text{A}}(n_j, n_j - 1) = 1, && n_j \geq 1,\\
&p_{js^-}^{\text{A}}(0, 0) = 1, &&\\
&q_{jc}^{\text{D}}(n_j, n_j - 1) = \mu_j, && n_j \geq 1.
\end{aligned}$$

The triggering probabilities are

$$\begin{aligned}
&f_{jc,u}(n_j, n_j') = 0, && u = c, s^+, s^- \text{ and } n_j \geq 0,\\
&f_{js^+,s^+}(n_j, n_j + 1) = 1, && n_j \geq 0,\\
&f_{js^-,s^-}(n_j + 1, n_j) = 1, && n_j \geq 0.
\end{aligned}$$

For convenience we let $\alpha_j = \alpha_{jc}$, $\alpha_j^+ = \alpha_{js^+}$ and $\alpha_j^- = \alpha_{js^-}$. Node j, characterized by

$$q_j^{(\alpha_j)}(n_j, n_j') = \alpha_j p_{jc}^{\text{A}}(n_j, n_j') + \alpha_j^+ p_{js^+}^{\text{A}}(n_j, n_j') + \alpha_j^- p_{js^-}^{\text{A}}(n_j, n_j') + q_{jc}^{\text{D}}(n_j, n_j'),$$

is an $M/M/1$ queue with Poisson arrivals of three classes of entities, with respective rates α_j, α_j^+ and α_j^-, and with service rate μ_j. As far as the stationary distribution of the node is concerned, this queue is the same as an $M/M/1$ queue with arrival rate $\alpha_j + \alpha_j^+$ and service rate $\mu_j + \alpha_j^-$. Hence its stationary distribution is

$$\pi_j(n_j) = (1 - \rho_j)\rho_j^{n_j}, \qquad n_j \geq 0,$$

provided the stability condition

$$\rho_j = (\alpha_j + \alpha_j^+)/(\mu + \alpha_j^-) < 1$$

is satisfied.

However, this node with positive and negative signals is not quasi-reversible, as we will see below. So we modify this queue by assuming that whenever it is empty, the node generates departures of class s^+ at a constant rate. That is, we modify the transition rate for node j such that $q^{\mathrm{D}}_{js^+}(0,0) > 0$. In what follows we show that this modified $M/M/1$ queue with positive and negative signals is quasi-reversible if and only if

$$q^{\mathrm{D}}_{js^+}(0,0) = \rho_j^{-1}\alpha_j^+. \tag{5.31}$$

It is easy to verify that the quasi-reversibility condition (5.21) is satisfied for $u = c, s^-$ with

$$\begin{aligned} \beta_j &= \rho_j \mu_j, \\ \beta_j^- (\equiv \beta_{js^-}) &= \rho_j \alpha_j^-. \end{aligned}$$

For $u = s^+$ and $n_j \geq 1$,

$$\begin{aligned} &\sum_{n_j'} \pi_j(n_j') \Big(q^{\mathrm{D}}_{js^+}(n_j', n_j) + \sum_{v=c,s^+,s^-} \alpha_{jv} p^{\mathrm{A}}_{jv}(n_j', n_j) f_{jv,s^+}(n_j', n_j) \Big) \\ &\qquad = \pi_j(n_j - 1)\alpha_j^+ p^{\mathrm{A}}_{js^+}(n_j - 1, n_j) \\ &\qquad = \rho_j^{-1} \alpha_j^+ \pi_j(n_j). \end{aligned}$$

And for $n_j = 0$,

$$\begin{aligned} &\sum_{n_j'} \pi_j(n_j') \Big(q^{\mathrm{D}}_{js^+}(n_j', 0) + \sum_{v=c,s^+,s^-} \alpha_{jv} p^{\mathrm{A}}_{jv}(n_j', 0) f_{jv,s^+}(n_j', 0) \Big) \\ &\qquad = q^{\mathrm{D}}_{js^+}(0,0)\pi_j(0). \end{aligned}$$

Thus for (5.21) to hold for $u = s^+$ and all n_j, (5.31) is necessary and sufficient. Letting

$$\beta_j^+ (\equiv \beta_{js^+}) = \rho_j^{-1} \alpha_j^+,$$

we obtain that node j is uniformly quasi-reversible for all α_j, α_j^+, α_j^- and

$$\begin{aligned} \beta_j &= \frac{\alpha_j + \alpha_j^+}{\mu_j + \alpha_j^-} \mu_j, \quad j = 1, \ldots, N, \\ \beta_j^- &= \frac{\alpha_j + \alpha_j^+}{\mu_j + \alpha_j^-} \alpha_j^-, \quad j = 1, \ldots, N, \\ \beta_j^+ &= \frac{\mu_j + \alpha_j^-}{\alpha_j + \alpha_j^+} \alpha_j^+, \quad j = 1, \ldots, N, \end{aligned}$$

provided $\rho_j < 1$ for all j. Hence, it follows from Theorem 3 that, if α_j, α_j^+ and α_j^- are the solutions of traffic equations (5.13), which are simplified to (5.27), (5.28) and (5.29), then the network has the geometric product form stationary distribution (5.30). This completes the proof of Theorem 4.

Since the modified network has additional departures of positive signals, the network process stochastically dominates the corresponding process without the additional departures if a positive signal cannot be transformed into a negative signal. This can be easily proved using sample path stochastic comparison by constructing the two processes and coupling the numbers of jobs in the two networks. Note that, if the modified network has a stationary distribution, and if it stochastically dominates the original network, then the original network must also have a stationary distribution. Thus we obtain the following result.

Corollary 2. Let π^0 be the stationary distribution of the network without the additional departure processes. If $r_{js^+,ks^-} = 0$ and $r_{jc,ks^-} = 0$ for all $j,k = 1,\ldots,N$, then π^0 is stochastically dominated by the product form geometric distribution obtained in Theorem 4, i.e.,

$$\sum_{k_j \ge n_j, j=1,\ldots,N} \pi^0(k_1,\ldots,k_N) \le \prod_{j=1}^{N} \left(\frac{\alpha_j + \alpha_j^+}{\mu_j + \alpha_j^-} \right)^{n_j}. \tag{5.32}$$

5.6.2 Multiple Classes of Positive and Negative Signals

We extend the results of the last subsection to networks with multiple classes of positive and negative signals. Suppose there is a single class of jobs denoted by c, I^+ classes of positive signals denoted by $\{u^+; u = 1,2,\ldots,I^+\}$, and I^- classes of negative signals denoted by $\{u^-; u = 1,2,\ldots,I^-\}$, where I^+ and I^- may be infinity. There is a single server at node j, and the processing times at node j are exponentially distributed with rate μ_j. Jobs arrive at node j from the outside according to a Poisson process with rate λ_j. Class u^+ positive signals, $u = 1,2,\ldots,I^+$, arrive at node j from the outside according to a Poisson process with rate λ_{ju}^+, and class u^- negative signals, $u = 1,2,\ldots,I^-$, arrive at node j from the outside according to a Poisson process with rate λ_{ju}^-.

The effects of positive and negative signals at a node are the same as before. That is, the arrival of a class u^+ positive signal at node j adds one job at node j and then departs, whereas the arrival of a class u^- negative signal at node j triggers one job, if any one is present, to depart. If a negative signal arrives at an empty node, nothing happens and the signal disappears. Thus, node j is characterized by

$$\begin{aligned}
p^{\mathrm{A}}_{jc}(n_j, n_j+1) &= 1, \qquad n_j \geq 0,\\
p^{\mathrm{A}}_{ju^+}(n_j, n_j+1) &= 1, \qquad n_j \geq 0, u = 1,2,\ldots,I^+,\\
p^{\mathrm{A}}_{ju^-}(n_j, n_j-1) &= 1, \qquad n_j \geq 1, u = 1,2,\ldots,I^-,\\
p^{\mathrm{A}}_{ju^-}(0,0) &= 1, \qquad u = 1,2,\ldots,I^-,\\
q^{\mathrm{D}}_{jc}(n_j, n_j-1) &= \mu_j, \qquad n_j \geq 1.
\end{aligned}$$

The triggering probabilities are

$$\begin{aligned}
f_{jc,w}(n_j, n_j') &= 0, \qquad w = c, u^+, v^-, \text{ and } n_j, n_j' \geq 0,\\
f_{ju^+,u^+}(n_j, n_j+1) &= 1, \qquad n_j \geq 0, u = 1,2,\ldots,I^+,\\
f_{ju^-,u^-}(n_j, n_j-1) &= 1, \qquad n_j > 0, u = 1,2,\ldots,I^-.
\end{aligned}$$

The routing probabilities are defined as follows. Upon a processing completion at node j, a job goes to node k as a regular job with probability $r_{jc,kc}$, as a class v^+ positive signal with probability r_{jc,kv^+}, as a class v^- negative signal with probability r_{jc,kv^-}, and it leaves the network with probability $r_{jc,0}$, where

$$\sum_{k=1}^{N}\left(r_{jc,kc} + \sum_{v=1}^{I^+} r_{jc,kv^+} + \sum_{v=1}^{I^-} r_{jc,kv^-}\right) + r_{jc,0} = 1, \qquad \text{for } j = 1,\ldots,N.$$

The arrival of a class u^+ positive signal at node j, either from the outside or from another node, adds one job to node j, and the signal leaves immediately for node k as a job with probability $r_{ju^+,kc}$, as a class v^+ positive signal with probability r_{ju^+,kv^+}, as a class v^- negative signal with probability r_{ju^+,kv^-}, and it leaves the network with probability $r_{ju^+,0}$, where

$$\sum_{k=1}^{N}\left(r_{ju^+,kc} + \sum_{v=1}^{I^+} r_{ju^+,kv^+} + \sum_{v=1}^{I^-} r_{ju^+,kv^-}\right) + r_{ju^+,0} = 1, \qquad \text{for } j = 1,\ldots,N, \text{ and } u = 1,2,\ldots,I^+.$$

Finally, the arrival of a class u^- negative signal at node j, either from the outside or from another node, triggers one job from the node to depart, provided the queue is not empty upon its arrival. The triggered job then goes to node k as a job with probability $r_{ju^-,kc}$, as a class v^+ positive signal with probability r_{ju^-,kv^+}, as a class v^- negative signal with probability r_{ju^-,kv^-}, and it leaves the network with probability $r_{ju^-,0}$, where

$$\sum_{k=1}^{N}\left(r_{ju^-,kc} + \sum_{v=1}^{I^+} r_{ju^-,kv^+} + \sum_{v=1}^{I^-} r_{ju^-,kv^-}\right) + r_{ju^-,0} = 1, \qquad \text{for } j = 1,\ldots,N, \text{ and } u = 1,2,\ldots,I^-.$$

The arrival of a negative signal at an empty node does not have any effect and disappears.

Remark 1. The only additional feature of the network with multiple classes of positive and negative signals is the class-dependent routing. However, the class-dependent routing is very useful and it is general enough to include most queueing networks with batch arrivals and batch processing as special cases. See Example 6.

Let α_{jc}, $\{\alpha^+_{ju}; j = 1,\ldots,N, u = 1,\ldots,I^+\}$, and $\{\alpha^-_{ju}; j = 1,\ldots,N, u = 1,\ldots,I^-\}$ denote the average arrival rates of jobs, positive signals, and negative signals at node j. They are determined by traffic equation (5.13), i.e.,

$$\alpha_j = \lambda_{jc} + \sum_{k=1}^{N} \rho_k \mu_k r_{kc,jc} + \sum_{k=1}^{N}\sum_{v=1}^{I^-} \rho_k \alpha^-_{kv} r_{kv^-,jc} + \sum_{k=1}^{N}\sum_{v=1}^{I^+} \rho_k^{-1} \alpha^+_{kv} r_{kv^+,jc}, \qquad j = 1,\ldots,N; \tag{5.33}$$

$$\alpha^+_{ju} = \lambda^+_{ju} + \sum_{k=1}^{N} \rho_k \mu_k r_{kc,ju^+} + \sum_{k=1}^{N}\sum_{v=1}^{I^-} \rho_k \alpha^-_{kv} r_{kv^-,ju^+} + \sum_{k=1}^{N}\sum_{v=1}^{I^+} \rho_k^{-1} \alpha^+_{kv} r_{kv^+,ju^+}, \qquad j = 1,\ldots,N, u = 1,\ldots,I^+ \tag{5.34}$$

$$\alpha^-_{ju} = \lambda^-_{ju} + \sum_{k=1}^{N} \rho_k \mu_k r_{kc,ju^-} + \sum_{k=1}^{N}\sum_{v=1}^{I^-} \rho_k \alpha^-_{kv} r_{kv^-,ju^-} + \sum_{k=1}^{N}\sum_{v=1}^{I^+} \rho_k^{-1} \alpha^+_{kv} r_{kv^+,ju^-}, \qquad j = 1,\ldots,N, u = 1,2,\ldots,I^- \tag{5.35}$$

where

$$\rho_j = \frac{\alpha_j + \alpha^+_j}{\mu_j + \alpha^-_j}, \qquad j = 1,\ldots,N,$$

and α^+_j and α^-_j are defined as

$$\alpha^+_j = \sum_{v=1}^{I^+} \alpha^+_{jv}, \qquad j = 1,\ldots,N,$$

$$\alpha^-_j = \sum_{v=1}^{I^-} \alpha^-_{jv}, \qquad j = 1,\ldots,N.$$

Clearly, α_j, α^+_j and α^-_j are the average arrival rates of jobs, positive signals and negative signals at node j. Also, the total average arrival rate of jobs, including regular job and those added by positive signals, is $\alpha_j + \alpha^+_j$.

Theorem 5. Suppose the traffic equations (5.33), (5.34) and (5.35) have nonnegative solutions such that

$$\rho_j \equiv \frac{\alpha_j + \alpha^+_j}{\mu_j + \alpha^-_j} < 1, \qquad \text{for all } j = 1,\ldots,N.$$

If the network is modified so that whenever node j is empty, there is an additional departure process of class u^+ positive signals with rate

$$\frac{\mu_j + \alpha_j^-}{\alpha_j + \alpha_j^+}\alpha_{ju}^+,$$

then the stationary probability of the network is

$$\pi(n) = \prod_{j=1}^{N} (1-\rho_j)\rho_j^{n_j}\,. \tag{5.36}$$

Corollary 3. Let π^0 be the stationary distribution of the network without the additional departures of positive signals. If $r_{ju^+,kv^-} = 0$ and $r_{jc,kv^-} = 0$ for all j,k, u and v, then π^0 is stochastically dominated by the geometric product form π of (5.36).

The following example illustrates how the multiple classes of signals can be used to model batch movements.

Example 6. Consider a network of N single-server nodes. Jobs arrive at node j from the outside according to a Poisson process with rate λ_j, $j = 1,2,\ldots,N$. The jobs are served in batches of a fixed size K_j, and the processing time of a batch is exponentially distributed with rate μ_j. Upon a processing completion at node j, the K_j jobs coalesce into a single job, and this single job goes to node k with probability r_{jk}, $k = 0,1,\ldots,N$, where 0 is the outside world. In case there are less than K_j jobs in node j upon a processing completion at the node, these jobs coalesce into a partial batch and are removed from the system. When a job arrives at node j when the number of jobs at the node is less than K_j, it joins the batch currently being served; otherwise it waits in queue.

This model is a special case of the network with multiple classes of negative signals. To see this, consider a network with a single class of jobs and $K_j - 1$ classes of negative signals at node j, denoted by u^- for $u = 1,2,\ldots,K_j - 1$. Jobs arrive at node j from the outside according to a Poisson process with rate λ_j. The jobs are served one at a time and the service rate at node j is μ_j. The routing probabilities, denoted by r^*, are defined as

$$\begin{aligned}
&r^*_{jc,j(K_j-1)^-} = 1, && j = 1,\ldots,N,\\
&r^*_{ju^-,j(u-1)^-} = 1, && u = 2,3,\ldots,K_j-1,\, j = 1,\ldots,N,\\
&r^*_{j1^-,kc} = r_{jk}, && k = 0,1,\ldots,N,\, j = 1,\ldots,N.
\end{aligned}$$

That is, upon a processing completion at node j a job goes back to node j as a class $(K_j - 1)^-$ negative signal with probability 1; the arrival of a class u^- $(u = 2,3,\ldots,K_j - 1)$ negative signal at node j removes one job and then immediately goes to node j as a class $(u-1)^-$ signal with probability 1; a class 1^- negative signal arrives at node j and reduces the number of jobs by 1, then goes to node k as a regular job with probability r_{jk}. This implies that a regular processing completion at node j instantaneously removes K_j jobs from node j, provided there are at least K_j jobs present, and then goes to node k as a regular job with probability r_{jk}. That a negative signal that arrives at an empty queue disappears translates to the fact that, when there are less than K_j jobs at node j upon a processing completion, the entire batch is removed from the network. This is exactly the network under consideration.

By Theorem 5, this network has a geometric product form stationary distribution. Since there are no positive signals, no additional departure process is required. □

5.7 Necessary and Sufficient Conditions for Product Form

In the preceding sections we discussed various network models that possess product form stationary distributions. Most of the results are obtained through quasi-reversibility. A natural question is whether quasi-reversibility is also a necessary condition for product form. The answer is negative. This section presents the necessary and sufficient conditions for product form for the class of networks whose transitions involve at most two nodes, i.e., there is no instantaneous triggering. Such a characterization yields a general procedure for verifying whether a network has a product form solution and obtaining it when it exists. Furthermore, the network has a product form stationary distribution and is *biased locally balanced* if and only if the network is quasi-reversible and certain traffic equations are satisfied. We also consider various scenarios in which quasi-reversibility is a necessary condition for product form.

The network consists of N nodes, indexed 1 to N, and the outside is labeled as node 0. However, unlike the formulation earlier sections, we here assume that the outside, i.e., node 0, has multiple states. Since departures from node 0 are arrivals to the network, such a formulation allows the arrival process to the network from the outside to be arbitrary. The state of the network is a vector of the states of the individual nodes and the outside world. Its state space is

$$\mathcal{S} = \mathcal{S}_0 \times \mathcal{S}_1 \times \cdots \times \mathcal{S}_N.$$

For convenience we only consider the case of single class of transitions. Extension multiple classes of transitions is straightforward. As in earlier sections, node j, $j = 0, 1, \ldots, N$, is subject to three types of state transitions, referred to as arrival, departure and internal transitions, denoted by

$$\{p_j^{\mathrm{A}}(x_j,y_j); x_j,y_j \in \mathcal{S}_j\},$$
$$\{q_j^{\mathrm{D}}(x_j,y_j); x_j,y_j \in \mathcal{S}_j\},$$
$$\{q_j^{\mathrm{I}}(x_j,y_j); x_j,y_j \in \mathcal{S}_j\}.$$

They represent, respectively, the transition probabilities due to arrivals, the transition rate due to departures, and the internal transition rate. Thus we must have

$$\sum_{y_j \in \mathcal{S}_j} p_j^{\mathrm{A}}(x_j, y_j) = 1, \qquad x_j \in \mathcal{S}_j.$$

The network process is characterized by the following system dynamics.

(i) When node j is in state x_j, the departure transition rate that changes the state from x_j to y_j is $q_j^{\mathrm{D}}(x_j, y_j), y_j \in \mathcal{S}_j$. .
(ii) A departure from node j is transferred to node k as an arrival with probability r_{jk}, $k = 0, 1, \ldots, N$ (recall that node 0 represents the outside).
(iii) An arrival at node k changes its state from x_k to y_k with probability $p_k^{\mathrm{A}}(x_k, y_k)$, $y_k \in \mathcal{S}_k$.
(iv) The internal transition rate at node j is $q_j^{\mathrm{I}}(x_j, y_j)$ when its state is x_j. We here redefine the internal transition so as to represent all transitions that do not trigger state changes at other nodes, i.e., it includes case (ii) with $j = k$. Denote this new internal transition rate by $q_j^{\mathrm{I}*}$, i.e.,

$$q_j^{\mathrm{I}*}(x_j, y_j) = q_j^{\mathrm{I}}(x_j, y_j) + \sum_{x_j'} q_j^{\mathrm{D}}(x_j, x_j') r_{jj} p_j^{\mathrm{A}}(x_j', y_j).$$

The network process has the transition rates

$$q(x, x') = \sum_{j,k} q_{jk}(x, x'), \qquad x, x' \in \mathcal{S},$$

where

$$q_{jk}(x, x') = \begin{cases} q_j^{\mathrm{D}}(x_j, x_j') r_{jk} p_k^{\mathrm{A}}(x_k, x_k') 1[y_\ell = x_\ell, \ell \neq j, k], & \text{if } j \neq k, \\ q_j^{\mathrm{I}*}(x_j, x_j') 1[x_\ell' = x_\ell, \ell \neq j], & \text{if } j = k. \end{cases}$$

Our objective is to find the necessary and sufficient conditions for the network to have a product form stationary distribution.

The following notation will be used in our analysis. For a probability distribution π_j on $\mathcal{S}_j$, define

$$\begin{aligned}
q_j^{\mathrm{D}}(x_j) &= \sum_{y_j} q_j^{\mathrm{D}}(x_j,y_j),\\
q_j^{\mathrm{I}*}(x_j) &= \sum_{y_j} q_j^{\mathrm{I}*}(x_j,y_j),\\
\tilde{p}_j^{\mathrm{A}}(x_j) &= \frac{\sum_{y_j} \pi_j(y_j)p_j^{\mathrm{A}}(y_j,x_j)}{\pi_j(x_j)},\\
\tilde{q}_j^{\mathrm{D}}(x_j) &= \frac{\sum_{y_j} \pi_j(y_j)q_j^{\mathrm{D}}(y_j,x_j)}{\pi_j(x_j)},\\
\tilde{q}_j^{\mathrm{I}*}(x_j) &= \frac{\sum_{y_j} \pi_j(y_j)q_j^{\mathrm{I}*}(y_j,x_j)}{\pi_j(x_j)},\\
\beta_j &= \sum_{x_j}\sum_{y_j} \pi_j(x_j)q_j^{\mathrm{D}}(x_j,y_j),\\
\nu_j &= \sum_{x_j}\sum_{y_j} \pi_j(x_j)q_j^{\mathrm{I}*}(x_j,y_j).
\end{aligned}$$

Notice that $q_j^{\mathrm{D}}(x_j)$ ($q_j^{\mathrm{I}*}(x_j)$) is different from the transition rate function $q_j^{\mathrm{D}}(x_j,y_j)$ ($q_j^{\mathrm{I}*}(x_j,y_j)$). They are distinguished only by their arguments. When they are used without arguments (e.g., q_j^{D}), they represent the transition rate functions, e.g., $q_j^{\mathrm{D}}(x_j,y_j)$. Assume that β_j and ν_j are finite. Keep in mind that $\tilde{p}_j^{\mathrm{A}}(x_j), \tilde{q}_j^{\mathrm{D}}(x_j), \tilde{q}_j^{\mathrm{I}*}(x_j)$ as well as β_j, ν_j are functions of π_j. The following relationships can be easily verified:

$$\sum_{x_j} \pi_j(x_j)q_j^{\mathrm{D}}(x_j) = \sum_{x_j} \pi_j(x_j)\tilde{q}_j^{\mathrm{D}}(x_j) = \beta_j, \tag{5.37}$$

$$\sum_{x_j} \pi_j(x_j)q_j^{\mathrm{I}*}(x_j) = \sum_{x_j} \pi_j(x_j)\tilde{q}_j^{\mathrm{I}*}(x_j) = \nu_j. \tag{5.38}$$

We first consider the possible forms of the marginal distributions when the network process has a product form stationary distribution. Define the transition rate q_j for each node j by

$$q_j(x_j,y_j) = \alpha_j p_j^{\mathrm{A}}(x_j,y_j) + (1-r_{jj})q_j^{\mathrm{D}}(x_j,y_j) + q_j^{\mathrm{I}*}(x_j,y_j), \quad x_j,y_j \in \mathcal{S}_j, \tag{5.39}$$

where α_j is a parameter to be determined. Consider this process as node j operating in isolation. The first term in the summation indicates that this isolated node has Poisson arrivals with rate α_j. The second and third terms are transition rates associated respectively with departures from node j and internal transitions at node j, where an internal transition may be a departure that returns to the same node (see (iv)).

Theorem 6. If the network process has the product form stationary distribution

$$\pi(x) = \prod_{j=0}^{N} \pi_j(x_j),$$

then each π_j is the stationary distribution for the q_j defined by (5.39) in which coefficients α_j are the solution to traffic equations

$$\alpha_j = \sum_{k \neq j} \beta_k(\alpha_k) r_{kj}, \qquad j = 0, 1, \ldots, N, \tag{5.40}$$

where $\beta_j(\alpha_j)$ denotes the β_j of (5.37) which depends on α_j through π_j. Note that in this, as well as the next, section we have included immediate feedback as internal transition, hence on the right hand side of the traffic equation only $k \neq j$ is needed.

Proof. The global balance equations for the network are

$$\pi(x) \sum_y q(x,y) = \sum_y \pi(y) q(y,x), \qquad x \in \mathcal{S}. \tag{5.41}$$

Since

$$\pi(y) = \frac{\pi(x)\pi_j(y_j)\pi_k(y_k)}{\pi_j(x_j)\pi_k(x_k)}, \tag{5.42}$$

for y such that $x_\ell = y_\ell$ for all $\ell \neq j,k$, it follows from the definition of q that (5.41) is equivalent to

$$\begin{aligned} &\pi(x) \sum_j \Big(q_j^{\mathrm{I}*}(x_j) + q_j^{\mathrm{D}}(x_j) \sum_{k \neq j} r_{jk} \Big) \\ &\qquad = \pi(x) \sum_j \Big(\tilde{q}_j^{\mathrm{I}*}(x_j) + \tilde{p}_j^{\mathrm{A}}(x_j) \sum_{k \neq j} r_{kj} \tilde{q}_k^{\mathrm{D}}(x_k) \Big), \qquad x \in \mathcal{S}. \end{aligned} \tag{5.43}$$

For a fixed j, we sum these equations over all x_ℓ for $\ell \neq j$. First, the left hand side becomes

$$\begin{aligned} &\sum_{x_\ell : \ell \neq j} \pi(x) \Big[q_j^{\mathrm{I}*}(x_j) + \sum_{j' \neq j} q_{j'}^{\mathrm{D}}(x_{j'}) r_{j'j} + q_j^{\mathrm{D}}(x_j) \sum_{k \neq j} r_{jk} \\ &\qquad + \sum_{j' \neq j} \Big(q_{j'}^{\mathrm{I}*}(x_{j'}) + q_{j'}^{\mathrm{D}}(x_{j'}) \sum_{k \neq j, j'} r_{j'k} \Big) \Big] \\ &= \pi_j(x_j) \Big(q_j^{\mathrm{I}*}(x_j) + \sum_{j' \neq j} \beta_{j'} r_{j'j} + (1 - r_{jj}) q_j^{\mathrm{D}}(x_j) \Big) + \sum_{j' \neq j} \Big(\nu_{j'} + \beta_{j'} \sum_{k \neq j, j'} r_{j'k} \Big) \\ &= \pi_j(x_j) \Big(q_j^{\mathrm{I}*}(x_j) + \alpha_j + (1 - r_{jj}) q_j^{\mathrm{D}}(x_j) \Big) + \sum_{j' \neq j} \Big(\nu_{j'} + \beta_{j'} \sum_{k \neq j, j'} r_{j'k} \Big). \end{aligned}$$

A similar manipulation on the right hand side yields

$$\pi_j(x_j) \Big(\tilde{q}_j^{\mathrm{I}*}(x_j) + \alpha_j \tilde{p}_j^{\mathrm{A}}(x_j) + (1 - r_{jj}) \tilde{q}_j^{\mathrm{D}}(x_j) \Big) + \sum_{j' \neq j} \Big(\nu_{j'} + \beta_{j'} \sum_{k \neq j, j'} r_{j'k} \Big).$$

Thus it follows from (5.43) that

$$q_j^{\mathrm{I}*}(x_j)+\alpha_j+(1-r_{jj})q_j^{\mathrm{D}}(x_j)=\tilde{q}_j^{\mathrm{I}*}(x_j)+\alpha_j\tilde{p}_j^{\mathrm{A}}(x_j)+(1-r_{jj})\tilde{q}_j^{\mathrm{D}}(x_j). \quad (5.44)$$

These are the balance equations for q_j divided by $\pi_j(x_j)$ with α_j given by (5.40). This completes the proof of Theorem 6.

The next theorem provides the necessary and sufficient conditions for the network process to have a product form distribution.

Theorem 7. The network has the product form stationary distribution

$$\pi(x)=\prod_{j=0}^{N}\pi_j(x_j), \qquad x\in\mathcal{S}$$

if and only if each π_j is the stationary distribution of q_j with coefficients α_j satisfying the traffic equations (5.40) and

$$(\tilde{q}_j^{\mathrm{D}}(x_j)-\beta_j)r_{jk}(\tilde{p}_k^{\mathrm{A}}(x_k)-1)+(\tilde{q}_k^{\mathrm{D}}(x_k)-\beta_k)r_{kj}(\tilde{p}_j^{\mathrm{A}}(x_j)-1)=0, \quad (5.45)$$

for all $j\neq k$ and $x_j\in\mathcal{S}_j$, $x_k\in\mathcal{S}_k$.

Proof. Assume the product form is $\pi(x)=\prod_{j=0}^{N}\pi_j(x_j)$. Since the conditions of Theorem 6 are satisfied, (5.43) holds. Dividing (5.43) by $\pi(x)$, and subtracting the summation of (5.44) over all j yields

$$\sum_j\Big(\alpha_j\tilde{p}_j^{\mathrm{A}}(x_j)+(1-r_{jj})\tilde{q}_j^{\mathrm{D}}(x_j)\Big)=\sum_j\Big(\alpha_j+\tilde{p}_j^{\mathrm{A}}(x_j)\sum_{k\neq j}r_{kj}\tilde{q}_k^{\mathrm{D}}(x_k)\Big). \quad (5.46)$$

For convenience define

$$D_{jk}(x_j,x_k)=(\tilde{q}_j^{\mathrm{D}}(x_j)-\beta_j)r_{jk}(\tilde{p}_k^{\mathrm{A}}(x_k)-1).$$

Since the stationary distribution is product form, it follows from Theorem 6 that α_j and β_j satisfy (5.40). Hence, substituting α_j of (5.40) into (5.46) gives

$$\sum_j\sum_{k\neq j}D_{jk}(x_j,x_k)=0. \quad (5.47)$$

Multiplying (5.47) by $\prod_{\ell\neq j,k}\pi_\ell(x_\ell)$, summing over x_ℓ for $\ell\neq j,k$, and observing that

$$\sum_{x_\ell}\pi_\ell(x_\ell)\left(D_{\ell k}(x_\ell,x_k)+D_{k\ell}(x_k,x_\ell)\right)=0$$

yields

$$D_{jk}(x_j,x_k)+D_{kj}(x_k,x_j)=0\,.$$

This is exactly (5.45).

Conversely, assume that each π_j is the stationary distribution of q_j and that (5.40) and (5.45) are satisfied. Since (5.45) implies (5.47), we obtain (5.46). Similarly, (5.44) follows from the fact that π_j is the stationary distribution of q_j. Hence, from the calculation of (5.46) we obtain (5.43). Thus the product form π satisfies the global balance (5.41).

Clearly, the following are three sufficient conditions for (5.45), so they are sufficient for the stationary distribution of the network to be product form.

(a) Both nodes j and k are quasi-reversible. Recall that node j is quasi-reversible if $\tilde{q}_j^{\mathrm{D}}(x_j)$ is independent of x_j; in this case it must equal to β_j.
(b) Both nodes j and k are non-effective with respect to arrivals. Node j is said to be *non-effective with respect to arrivals* if $\tilde{p}_j^{\mathrm{A}}(x_j) = 1$ for all $x_j \in \mathcal{S}_j$.
(c) Either node j or node k is quasi-reversible and non-effective with respect to arrivals.

These sufficient conditions are further weakened if $r_{jk} = 0$ or $r_{kj} = 0$. These and other special cases will be discussed in the next section. Note that when the outside (node 0) is a Poisson source, then node 0 has only one state (say 0) which is non-effective with respect to arrivals, i.e., the state of the outside source is not changed when a job departs the network. On the other hand, the Poisson source is clearly quasi-reversible, so it belongs to case (c). Therefore, when a network is subject to Poisson arrivals from the outside, condition (5.45) only has to be verified for nodes other than 0. In this case the product form stationary distribution $\prod_{j=0}^{N} \pi_j(x_j)$ can be written as $\prod_{j=1}^{N} \pi_j(x_j)$.

Theorem 7 yields the following procedure for establishing the existence of a product form stationary distribution for the network process and obtaining the distribution when it exists.

Step 1. For the dummy parameter α_j compute the stationary distribution π_j of node j defined by q_j of (5.39).
Step 2. Compute β_j using (5.40), which is a function of α_j since π_j is. So write it as $\beta_j(\alpha_j)$.
Step 3. Solve the traffic equations (5.40).
Step 4. Check condition (5.45) for each pair j,k and all x_j,x_k.

If this four-step procedure is successful, then $\pi(x) = \prod_{j=0}^{N} \pi_j(x_j)$ is the stationary distribution of the network process.

Finding vector $\alpha = (\alpha_0, \alpha_1, \dots, \alpha_N)$ that satisfies the traffic equations (5.40) is a fixed point problem whose solution is usually established by Brouwer's fixed point theorem. It follows from Theorem 7 that such a fixed point always exists when the network has a product form stationary distribution.

Theorem 8. The network has a product form stationary distribution if and only if there exists a solution to the traffic equations (5.40) and it satisfies the condition of Step 4.

Therefore, the procedure above, in principle, applies to any queueing networks with product solutions. If the procedure is successful it gives the product form stationary distribution of the network; otherwise, i.e., if it does not lead to a solution that satisfies the condition of Step 4, then the procedure concludes that the network does not has a product form solution. For a particular application, one may be able to construct an algorithm to compute the fixed point, e.g., an iterative algorithm.

5.8 Quasi-Reversibility Revisited

As we observed earlier, quasi-reversibility is a sufficient condition but not a necessary condition for product form. This section is concerned with how much stronger than necessary this condition is. It turns out that quasi-reversibility is equivalent to a product form that satisfies the biased local balance equations.

A Markov chain with transition rate q is said to satisfy *biased local balance* with respect to a positive probability measure π on $\mathcal{S}$ and real numbers $\gamma = \{\gamma_j; j = 0,1,\dots,N\}$ if $\sum_j \gamma_j = 0$ and

$$\pi(x)\left(\sum_k \sum_y q_{jk}(x,y) + \gamma_j\right) = \sum_k \sum_y \pi(y) q_{kj}(y,x), \quad x \in \mathcal{S},\ j = 0,1,\dots,N. \tag{5.48}$$

The π must be the stationary distribution for the Markov chain q since the global balance equations are the sum of these biased local balance equations over j. Also, we say that q is *locally balanced* with respect to π when all the γ_j's are 0.

Theorem 9. The following statements are equivalent.

(i) The network satisfies biased local balance with respect to a product form distribution $\pi(x) = \prod_{j=0}^N \pi_j(x_j)$ and $\gamma = \{\gamma_j; j = 0,1,\dots,N\}$.
(ii) Each node q_j is quasi-reversible with respect to π_j for some α_j that satisfies

$$\alpha_j = \sum_{k \neq j} \beta_k r_{kj}, \quad j = 0,1,\dots,N. \tag{5.49}$$

If these statements hold, then

$$\gamma_j = \alpha_j - (1 - r_{jj})\beta_j, \quad j = 0,1,\dots,N. \tag{5.50}$$

Proof. Suppose (i) holds. Since π has the product form, it follows from Theorem 6 that π_j is the stationary distribution of q_j. Using the same argument that we derived (5.44) from (5.42), we obtain the following equation from the biased local balance equation (5.48):

$$q_j^{\mathrm{I}*}(x_j) + (1 - r_{jj}) q_j^{\mathrm{D}}(x_j) + \gamma_j = \tilde{q}_j^{\mathrm{I}*}(x_j) + \tilde{p}_j^{\mathrm{A}}(x_j) \sum_{k \neq j} \tilde{q}_k^{\mathrm{D}}(x_k) r_{kj}. \tag{5.51}$$

Define α_j by (5.49) and fix node j. Multiplying (5.51) by $\pi_j(x_j)$, summing over x_j and applying (5.37) and (5.38), we obtain

$$(1 - r_{jj})\beta_j + \gamma_j = \sum_{k \neq j} \tilde{q}_k^{\mathrm{D}}(x_k) r_{kj}.$$

Fix $\ell \neq j$. Multiplying this equation by $\prod_{k\neq j,\ell} \pi_k(x_k)$, summing over x_k for $k \neq j,\ell$ and applying (5.37) and (5.49) yields

$$\begin{aligned}(1-r_{jj})\beta_j + \gamma_j &= \tilde{q}^{\mathrm{D}}_{\ell}(x_\ell) r_{\ell j} 1(\ell \neq j) + \sum_{k\neq j,\ell} \beta_k r_{kj} \\ &= \alpha_j + (\tilde{q}^{\mathrm{D}}_{\ell}(x_\ell) - \beta_\ell) r_{\ell j} 1(\ell \neq j) \end{aligned} \tag{5.52}$$

Summing over j and using (5.49) gives rise to

$$(\tilde{q}^{\mathrm{D}}_{\ell}(x_\ell) - \beta_\ell) \sum_{j\neq \ell} r_{\ell j} = 0.$$

This proves $\tilde{q}^{\mathrm{D}}_{\ell}(x_\ell) = \beta_\ell$. Thus, each node q_ℓ is quasi-reversible, so (ii) is proved.

Next assume (ii) holds. Since quasi-reversibility implies (5.45), the conditions for the product form of Theorem 7 are satisfied, and therefore (5.44) holds. Substituting $p^{\mathrm{A}}_j(x_j) = 1$ and $\tilde{q}^{\mathrm{D}}_j(x_j) = \beta_j$ in (5.44), we obtain

$$q^{\mathrm{I}*}_j(x_j) + \alpha_j + (1-r_{jj}) q^{\mathrm{D}}_j(x_j) = \tilde{q}^{\mathrm{I}*}_j(x_j) + \alpha_j \tilde{p}^{\mathrm{A}}_j(x_j) + (1-r_{jj})\beta_j, \quad x_j \in \mathcal{S}_j.$$

Define γ_j by (5.50). Applying (5.50) to the expression above yields

$$q^{\mathrm{I}*}_j(x_j) + (1-r_{jj}) q^{\mathrm{D}}_j(x_j) + \gamma_j = \tilde{q}^{\mathrm{I}*}_j(x_j) + \alpha_j \tilde{p}^{\mathrm{A}}_j(x_j), \quad x_j \in \mathcal{S}_j. \tag{5.53}$$

From (5.49) and the fact that $\beta_j = \tilde{q}^{\mathrm{D}}_j(x_j)$, it follows that

$$\alpha_j = \sum_{k\neq j} \tilde{q}^{\mathrm{D}}_k(x_k) r_{kj}.$$

Substituting this α_j into (5.53) yields (5.51), which implies (5.48). Hence q satisfies biased local balance with respect to π and γ. This completes the proof that (ii) implies (i).

The remaining part of this section explores scenarios under which quasi-reversibility is also a necessary condition for product form.

Corollary 4. If in a queueing network there are no immediate turn around loop, i.e., $r_{jk} \neq 0$ implies $r_{kj} = 0$, and $p^{\mathrm{A}}_k(x_k)$ is not identically 1 for all x_k, i.e.,

$$\tilde{p}^{\mathrm{A}}_k(x_k) \neq 1, \quad \text{for at least one } x_k \in \mathcal{S}_k, \tag{5.54}$$

then the product form implies that node j is quasi-reversible. In particular, if the discrete-time Markov chain on $\mathcal{S}_k$ with transition probability $\{p^{\mathrm{A}}_k(x_k,y_k); x_k,y_k \in \mathcal{S}_k\}$ is transient, then (5.54) is satisfied.

Proof. Under the assumptions, equation (5.45) is reduced to

$$(\tilde{q}^{\mathrm{D}}_j(x_j) - \beta_j) r_{jk} (\tilde{p}^{\mathrm{A}}_k(x_k) - 1) = 0. \tag{5.55}$$

Fix an x_k that satisfies (5.54). Since (5.55) holds for every x_j and $r_{jk}(\tilde{p}_k^{\mathrm{A}}(x_k)-1)\neq 0$, we conclude that $\tilde{q}_j^{\mathrm{D}}(x_j)=\beta_j$ for all x_j, i.e., node j is quasi-reversible.

To prove the second part, first note that $\tilde{p}_k^{\mathrm{A}}(x_k)=1$ for all x_k is, by the definition of $\tilde{p}_k^{\mathrm{A}}(x_k)$, equivalent to π_k being a positive stationary measure for the Markov chain with transition probability $\{p_k^{\mathrm{A}}(x_k,y_k);x_k,y_k\in\mathcal{S}_k\}$; this cannot be true if p_k^{A} is transient. Thus there must be at least one x_k such that (5.54) is satisfied.

In job based queues with no signals, arrivals do not decrease the number of jobs at the node, so p_k^{A} is clearly transient. This is not true, however, in networks with negative signals, in which the arrival of a negative signal reduces the number of regular jobs.

Definition 5. Node j is said to be *non-terminal* if

$$1-r_{jj}-r_{j0}>0\,,$$

i.e., a departure from node j arrives at other nodes in the network with a positive probability.

Feedforward networks clearly satisfy the first condition of Corollary 4 for all nodes that are non-terminal. Thus we obtain the following result.

Corollary 5. A job based feedforward queueing network has a product form stationary distribution if and only if all the non-terminal nodes are quasi-reversible.

Of course, there are many queueing networks with feedback that satisfy the conditions of Corollary 4.

Example 7. Consider the job based network with four nodes. Jobs arrive at nodes 1 and 2 according to Poisson processes. Departures from nodes 1 and 2 join node 3, and departures from node 4 either join node 1, node 2, or leave the network. Clearly, this network satisfies the conditions of Corollary 5, so quasi-reversibility of each node is both necessary and sufficient for the stationary distribution of the network to be product form.

To present the next result, we need to introduce two new concepts. Node j is called a *conventional queue* if it has an empty state, denoted by 0, from which there can be no departures or internal transitions, and state 0 cannot be reached via arrival or internal transitions. That is,

$$\begin{aligned}&\tilde{p}_j^{\mathrm{A}}(0)=q_j^{\mathrm{D}}(0)=0,\\&q_j^{\mathrm{I}*}(x_j,x_j')=0,\qquad\text{if either } x_j=0 \text{ or } x_j'=0.\end{aligned}$$

Clearly, if a network has an outside Poisson source, then node 0 is not conventional. A queueing network is called conventional if its outside source is Poisson and all other nodes are conventional. Let α_j^+ denote the average arrival rate at node j including the feedback, i.e.,

$$\alpha_j^+ = \sum_{k=0}^{N} \beta_k r_{kj} = \alpha_j + r_{jj}\beta_j \,.$$

Node j is said to be *internally balanced*, if $\alpha_j^+ = \beta_j$, i.e., the average arrival rate equals the average departure rate.

Theorem 10. Suppose a queueing network has conventional nodes and the outside source, i.e., node 0, is non-effective with respect to arrivals. If the network has a product form stationary distribution, then a non-terminal node k is quasi-reversible if and only if either one of the following two conditions holds.

(a)Node k has a path connecting it to an internally balanced node.
(b)Node k is directly connected to some node $j \neq 0$, but node j is not directly connected to node k, i.e, $r_{kj} > 0$ and $r_{jk} = 0$.

In these cases, all the non-terminal nodes are internally balanced. Note that condition (b) is satisfied if the destination node k is a terminal node.

Proof. Suppose the network has a product form stationary distribution. By Theorem 6, the marginal distribution for each node satisfies the balance equation (5.44) for the coefficients determined by the traffic equations (5.40). Substituting $x_j = 0$ in (5.44) yields

$$\alpha_j + (1-r_{jj})q_j^{\mathrm{D}}(0) + q_j^{\mathrm{I}*}(0) = \alpha_j \tilde{p}_j^{\mathrm{A}}(0) + (1-r_{jj})\tilde{q}_j^{\mathrm{D}}(0) + \tilde{q}_j^{\mathrm{I}*}(0)\,.$$

Thus it follows from the condition that node j is a conventional queue that

$$\alpha_j = (1-r_{jj})\tilde{q}_j^{\mathrm{D}}(0)\,. \tag{5.56}$$

Because of the product form, we also have (5.45) of Theorem 7. Letting $x_j = 0$ in (5.45) and substituting (5.56), we obtain

$$(\alpha_j^+ - \beta_j)\frac{r_{jk}}{1-r_{jj}}(\tilde{p}_k^{\mathrm{A}}(x_k) - 1) - (\tilde{q}_k^{\mathrm{D}}(x_k) - \beta_k)r_{kj} = 0, \quad k \neq j, 0. \tag{5.57}$$

Substituting $x_k = 0$ in (5.57) yields

$$(\alpha_j^+ - \beta_j)\frac{r_{jk}}{1-r_{jj}} + (\alpha_k^+ - \beta_k)\frac{r_{kj}}{1-r_{kk}} = 0. \tag{5.58}$$

Since the network process is irreducible, any non-terminal node k has an arc directly connecting it to some other node j, i.e., $r_{kj} > 0$. From (5.57) it follows that node k is quasi-reversible if and only if

$$(\alpha_j^+ - \beta_j)r_{jk}(\tilde{p}_k^{\mathrm{A}}(x_k) - 1) = 0\,.$$

Letting $x_k = 0$ in this formula we conclude either $\alpha_j^+ - \beta_j = 0$ or $r_{jk} = 0$. The latter is exactly (b), while the former is (a) since node j is internally balanced. Thus (a) and (b) are necessary for the quasi-reversibility of node k. Conversely, from (5.57),

(b) clearly implies that node k is quasi-reversible; if (a) is satisfied, we need to use induction to show that node k is quasi-reversible. First, if node k is directly connected to node j, i.e., $r_{kj} > 0$, then it follows from (5.57) that node k is quasi-reversible. If node k is connected to j through $i_1, i_2, \ldots, i_\ell$, then applying (5.58) to nodes i_ℓ and j shows that node i_ℓ is internally balanced, and applying (5.58) to nodes $i_{\ell-1}$ and i_ℓ shows that node $i_{\ell-1}$ is internally balanced, etc. After showing that node i_1 is internally balanced, we apply (5.57) to nodes k and i_1 to obtain that node k is quasi-reversible.

To show that each non-terminal node has to be internally balanced, assume that node j is quasi-reversible. Then, from (5.56) and $\tilde{q}_j^{\mathrm{D}}(0) = \beta_j$ it follows that

$$\alpha_j^+ = \alpha_j + r_{jj}\beta_j = \alpha_j + r_{jj}\tilde{q}_j^{\mathrm{D}}(0) = \tilde{q}_j^{\mathrm{D}}(0) = \beta_j\,.$$

The proof of Theorem 10 is thus complete.

From sufficient condition (c) of Section 7 for product form stationary distribution, it follows that if the outside source is Poisson, then no condition is required on the terminal nodes for the product form to hold.

The next result follows immediately from Theorem 10.

Corollary 6. Consider a queueing network with all nodes being conventional and all non-terminal nodes satisfy either (a) or (b) of Theorem 10. The network has a product form stationary distribution if and only if all the non-terminal nodes are quasi-reversible.

Considering an even more special case we obtain the following result.

Corollary 7. Consider a conventional queueing network with Poisson arrivals from the outside and each node is internally balanced, i.e., the departure rate of each node is equal to its arrival rate. The network has a product form stationary distribution if and only if all non-terminal nodes are quasi-reversible. If the outside source node is also a conventional queue, then the network has a product form stationary distribution if and only if all nodes are quasi-reversible.

Note the difference between Corollary 5 and Corollary 7. In Corollary 6 the non-terminal nodes do not need to be conventional or internally balanced, but the topological structure of the network is restricted. On the other hand, in Corollary 7, the network topology may be arbitrary, but each non-terminal node has to be conventional and internally balanced.

5.9 Networks with Random Customer Shuffling

In this section we show that the results on product form networks can be extended to models with random reshuffling of customer positions at both customer/signal arrival and departure epochs. Random permutation of customers at each node has been studied by several authors, including Yashkov (1980), Daduna (2001), Daduna and Schassberger (1985), Yates (1994), and Bonald and Tran (2007). The models studied in Yashkov (1980), Daduna (2001), Daduna and Schassberger (1983) and Yates (1994) are discrete time, while Bonald and Tran (2007) consider continuous time model. Though most of the networks discussed earlier in this chapter can be extended to include this additional feature of random shuffling, for simplicity in what follows we consider the case with multiple classes of customers but only one class of negative signals.

The network has N nodes and I classes of customers and one class of negative signals. Class u customers arrive to node i from the outside according to a Poisson process with rate λ_{iu}, and each class u customer requires exponentially distributed amount of time at node i with mean $1/\mu_{iu}$. Let n_{iu} be the number of class u customers at node i, and

$$\begin{aligned} n_i &= (n_{i1},\ldots,n_{iI}),\\ n &= (n_1,\ldots,n_N). \end{aligned}$$

Suppose n_i is the total number of customers at node i, i.e.,

$$n_i = \sum_{u=1}^{I} n_{iu}.$$

Let e_{iu} represent the unit vector with a 1 for class u customer at node i and 0 elsewhere.

We shall refer to $n = (n_1,\ldots,n_N)$ as the macro-state of the network. The micro-state, c to be defined below, shall include information about the classes of customers and their positions at the nodes.

Since we consider multiple classes of customers, the service disciplines at the node is assumed to be symmetric. However, we shall assume that the service discipline at node i depends not only on the number of customers at node i but also on the macro-state of the entire network. When the network is in macro-state n, let $\gamma_i(\ell,n)$ be the proportion of service effort at node i that is directed to position ℓ, $\ell = 1,\ldots,n_i$. Similarly, when a customer arrives at node i, it joins position ℓ with probability $\gamma_i(\ell,n)$, and customers originally at station $\ell,\ell+1,\ldots,n_i$ move to $\ell+1,\ell+2,\ldots,n_i+1$, respectively. When a class u customer finishes service at node i, it leaves for node j as a class v customer with probability $r_{iu,jv}$, it leaves for node j as a signal with probability $r_{iu,js}$, and it departs the network with probability $r_{iu,0}$, where

$$\sum_{j=1}^{N}\sum_{v=1}^{I} r_{iu,jv} + \sum_{j=1}^{N} r_{iu,js} + r_{iu,0} = 0, \qquad i = 1,\ldots,N, u = 1,\ldots,I.$$

Negative signals, denoted by s, arrive at node i from the outside according to a Poisson process with rate λ_i^-. When a signal arrives at node i, it heads for position ℓ and deletes the customer in position ℓ with probability $\gamma_i(\ell,n)$.

Given that there are n_i customers in node i, let c_i be the state of node i which is the sequence of n_i elements whose ℓ-th value $c_i(\ell)$ is the class of the customer in position ℓ, i.e.,

$$c_i = (c_i(1), c_i(2), \ldots, c_i(n_i)).$$

Let

$$c = (c_1, \ldots, c_N)$$

be the state of the network, which is referred to as the micro-state.

Let α_{iu} be the overall arrival rate of class u customers at node i, and α_i^- the overall arrival rate of signals at node i. Then the following traffic equations are satisfied:

$$\begin{aligned} \alpha_{iu} &= \lambda_{iu} + \sum_{k=1}^{N}\sum_{v=1}^{I} \frac{\alpha_{kv}\mu_{kv}}{\mu_{kv}+\alpha_k^-} r_{kv,iu}, \quad i = 1, \ldots, N, u = 1, \ldots, I, \\ \alpha_i^- &= \lambda_i^- + \sum_{k=1}^{N}\sum_{v=1}^{I} \frac{\alpha_{kv}\alpha_k^-}{\mu_{kv}+\alpha_k^-} r_{kv,is}, \quad i = 1, \ldots, N. \end{aligned}$$

The additional feature for the network is *random shuffling*. That is, immediately *after* the arrival of a customer or a signal, and immediately *after* a departure from a node, the customers at each and every node randomly permute positions within the same node, and this happens at all nodes simultaneously. The results of this section also hold true when the customers are assumed to randomly shuffle positions immediately *before* the arrival of customers or signals. But in this section we shall focus on the case that the shuffling takes place immediately after arrival and immediately after departure.

Let $\mathcal{P}(m)$ denote the set of permutations of m elements, $m \geq 1$, and let $\mathcal{P}(0)$ denote the identity mapping on $\{\emptyset\}$. Let

$$\mathcal{P}(n) = \mathcal{P}(n_1) \times \mathcal{P}(n_2) \times \cdots \times \mathcal{P}(n_N).$$

For any micro-state c and any permutation $\sigma \in \mathcal{P}(n)$, we let $\sigma(c)$ denote the micro-state whose i-th component is equal to $\sigma_i(c_i)$, i.e., the positions of customers in node i are permuted according to $\sigma_i \in \mathcal{P}_i(n_i)$.

For a macro-state n, let $\alpha_{iu}^A(\cdot, n), \alpha_{is}^A(\cdot, n), \alpha_{iu}^D(\cdot, n)$ be arbitrarily given distributions on $\mathcal{P}(n)$, and they are interpreted as follows: Immediately after a class u customer arrives at node i, all customers in the network randomly shuffle positions according distribution $\alpha_{iu}^A(\cdot, n)$, where n is the macro-state of the network *after* the class u customer joins node i; immediately after a negative signal arrives at node i, the customers in the network randomly shuffle positions according to permutation probability $\alpha_{is}^A(\cdot, n)$, where n is the state after the arrival of the signal; and immediately after a class u departs from node i, and before its arrival at the destination node, all the customers in the network randomly shuffle according to probability $\alpha_{iu}^D(\cdot, n)$, here again n is the macro-state after the departure of the customer but before its arrival at another node. We note here that a departure from a node causes the customers in the network to reshuffle twice (unless the departure is heading for the outside of the network): the first shuffle takes place after the departure, while the second happens after it arrives at the destination node.

The following result shows that, the introduction of the additional feature of random shuffling does not affect the stationary distribution of the network.

Theorem 11. If the solution to traffic equations satisfy $\sum_{u=1}^{I}\alpha_{iu} < \mu_{iu}+\alpha_i^-$ for $i=1,\dots,N$ and $u=1,\dots,I$, then the stationary distribution of the network with random shuffling is

$$\pi(c) = \prod_{i=1}^{N}\pi_i(c_i),$$

$$\pi(n) = \prod_{i=1}^{N}\pi_i(n_i),$$

where, for $i=1,\dots,N$,

$$\pi_i(c_i) = \left(1-\sum_{u=1}^{I}\frac{\alpha_{iu}}{\mu_{iu}+\alpha_i^-}\right)\prod_{\ell=1}^{n_i}\frac{\alpha_{ic_i(\ell)}}{\mu_{ic_i(\ell)}+\alpha_i^-},$$

$$\pi_i(n_i) = \left(1-\sum_{u=1}^{I}\frac{\alpha_{iu}}{\mu_{iu}+\alpha_i^-}\right)\frac{n_i!}{\prod_{u=1}^{I}n_{iu}!}\prod_{u=1}^{I}\left(\frac{\alpha_{iu}}{\mu_{iu}+\alpha_i^-}\right)^{n_{iu}}.$$

Proof. We show that the stationary distribution satisfies the cross balance equations for each class of customer u in each position ℓ of the associated node i (Chao, Miyazawa, and Pinedo (1999)). Notice that the distribution $\pi(c)$ only depends on the classes of customers at each node, and it does not depend on their relative positions at the node. This, turns out to be the key for the network distribution to be not affected by customer shuffling within the node.

First, we introduce some notation. For any network micro-state c, let $c\oplus e_{iu}(\ell)$ be the state after a class u customer joins position ℓ of node i, and let $c\ominus e_{ic_i(\ell)}(\ell)$ be the state of the network after the customer at position ℓ of node i leaves the system. For convenience let $\rho_{iu}=\alpha_{iu}^+/(\mu_{iu}+\alpha_i^-)$.

Consider a micro-state c such that $c_i(\ell)=u$. The macro-state is n. The probability flux corresponding to a departure in position ℓ from node i in micro-state c, either due to service completion or removal by signals, causing the network state to go across c from above, is

$$\begin{aligned}
&\pi(c)\mu_{iu}\gamma_i(\ell,n)+\pi(c)\lambda_i^-\gamma_i(\ell,n)\\
&\quad+\sum_{j=1}^{N}\sum_{\ell=1}^{n_j+1}\sum_{\sigma\in\mathcal{P}(n)}\pi(c')\mu_{jv}\gamma_j(\ell,n+e_{jv})r_{jv,is}\delta_{jv}^{D}(\sigma,n)\gamma_i(\ell,n)\\
&=\pi(c)\Big(\mu_{iu}+\lambda_i^-+\sum_{j=1}^{N}\sum_{v=1}^{I}\rho_{jv}\mu_{jv}r_{jv,is}\Big)\gamma_i(\ell,n)\\
&=\pi(c)(\mu_{iu}+\alpha_i^-)\gamma_i(\ell,n),
\end{aligned}$$

where c' is such that $c'_i(\ell) = v$ and $\sigma(c' \ominus e_{jv}(\ell)) = c$, and the last equality follows from the traffic equation. Similarly, the probability flux corresponding to an arrival of class u customer at position ℓ of node i, causing the state of the network to go across state c from below, is

$$\begin{aligned}
&\sum_{\sigma \in \mathcal{P}(n)} \sum_{\ell=1}^{n_i} \pi(c')\lambda_{iu}\delta^A_{iu}(\sigma,n)\gamma_i(\ell,n) \\
&+\sum_{j=1}^{N} \sum_{\ell'=1}^{n_j+1} \sum_{\sigma' \in \mathcal{P}(n-e_{iu})} \sum_{\sigma'' \in \mathcal{P}(n)} \pi(c'')\mu_{jv}\gamma_j(\ell',n+e_{jv}-e_{iu})r_{jv,iu}\delta^D_{jv}(\sigma',n-e_{iu})\delta^D_{iu}(\sigma'',n)\gamma_j(\ell,n) \\
&= \pi(c)\frac{\lambda_j + \sum_{j=1}^{N}\sum_{u=1}^{I} \rho_{jv}\mu_{jv}r_{jv,iu}}{\rho_{iu}}\gamma_i(\ell,n) \\
&= \pi(c)\alpha^+_{iu}/\rho_{iu}\gamma_i(\ell,n) \\
&= \pi(c)(\mu_{iu}+\alpha^-_i)\gamma_i(\ell,n),
\end{aligned}$$

where the second equality follows from the traffic equation, and the last equality follows from the definition of ρ_{iu}, and c' and c'' are such that

$$\begin{aligned}
\sigma(c' \oplus e_{iu}(\ell)) &= c, \\
\sigma''\Big(\sigma'(c'' \ominus e_{jv}(\ell')) \oplus e_{iu}(\ell)\Big) &= c.
\end{aligned}$$

This shows that the cross balance equations is satisfied for any position ℓ of any node i with regard to any class of customers, implying that the global balance equations are satisfied. This proves that $\pi(c)$ is the stationary distribution of the network. The stationary probability $\pi(n)$ is implied by $\pi(c)$.

Remark 2. All the results in this section holds true after introducing multiple classes of negative signals.

Remark 3. As mentioned earlier, the results remain the same when customers randomly shuffle positions immediately before the arrivals of customers and/or signals.

Remark 4. Haviv (2005) considers a "one-chance random queue", where an arrival always joins the head of the queue, however at a service completion a customer is randomly selected for service. The "one-chance random queue" is clearly a special case of random shuffling of this section, since that service discipline can be obtained by the LIFO service discipline followed by random reshuffling of customers immediately after a service completion.

5.10 Conclusion

In this chapter we reviewed some latest developments on queueing networks with tractable stationary distributions. Clearly, if possible it is always preferred to find the closed form analytical solution for a network problem, and only when this is not possible will one resort to approximation methods. Furthermore, we note that even when analytical solution is not available, the necessary and sufficient condition often can help obtain bounds and approximations for the non-product form networks. This is because of the fact that necessary and sufficient condition reveals the additional conditions that need to be imposed for the network problem to yield a product form solution. In many cases, the network after imposing additional conditions, that has a product form solution, gives rise to a stochastic bound for the original problem. Moreover, it is clear that, if the additional conditions only have minor impact on the performance of the original problem, then the product form solution obtained can be used as a good approximation for the original problem.

This chapter focused on queueing network models with exponential processing times. For models with arbitrary processing time distributions, state-dependent transition rates (such as multi-server queues, etc.), and discrete time models, the reader is referred to Chao, Miyazawa and Pinedo (1999) and the other papers in the references.

References

1. N. Asaddathorn and X. Chao (1999), "A decomposition approximation for assembly-disassembly types of queueing networks", *Annals of Operations Research*, 87, 247-261, 1999.
2. F. Baskett, K. M. Chandy, R. R. Muntz and F. G. Palacios (1975), "Open, closed and mixed networks of queues with different classes of customers," *J of ACM*, 22, 248-260.
3. Bonald, T. and M-A. Tran (2007). "On Kelly networks with shuffling". Working paper, Ecole Normale Superieure, France.
4. R. Boucherie and X. Chao (2001), "Queueing networks with string transitions of mixed vector additions and vector removals," *J. Systems Science and Complexity*, 14, 337-355.
5. R. Boucherie, X. Chao, and M. Miyazawa (2003), "Arrival first queueing networks with applications in Kanban production systems," *Performance Evaluation*, 51, 83-102.
6. Chandy, K. M., Howard, J. H. Jr. and Towsley, D. F. (1977), "Product form and local balance in queueing networks," *J of ACM* 24, 250-263.
7. X. Chao, (1994), "A Note On Networks of Queues with Signals and Random Triggering Times", *Probability in the Engineering and Informational Sciences*, Vol. 8, 213-219.
8. X. Chao, (1995),"On Networks of Queues with Arbitrary Processing Time Distributions", *Operations Research*, 43, 537-544.
9. X. Chao, (1995), "A Queueing Network Model with Catastrophe and Product Form Solution", *Operations Research Letters*, 18, 75-79.
10. X. Chao, (1997), "Partial balances in batch arrival, batch service and assemble-transfer queueing networks", *Journal of Applied Probability*, 34, 745-752.
11. X. Chao, W. Henderson and P. Taylor, (2001), "State-dependent coupling of general queueing networks", *Queueing Systems: Theory and Application*, 39, 337-348.
12. X. Chao and M. Miyazawa, (1998) "On quasi-reversibility and partial balance: An unified approach to product form results", *Operations Research*, Vol 46, 927-933.
13. X. Chao and M. Miyazawa, (2000), "Queueing networks with instantaneous movements: A unified approach by quasi-reversibility", *Advances in Applied Probability*, 32, 284-313.
14. X. Chao and M. Miyazawa, (2000), "On truncation properties for finite buffer queues and queueing networks", *Probability in the Engineering and Informational Sciences*, 14, 409-423.
15. X. Chao, M. Miyazawa, and M. Pinedo, (1999), *Queueing Networks: Customers, Signals, and Product Form Solutions*, John Wiley & Sons, Chichester.

16. X. Chao, M. Miyazawa, R. Serfozo, and H. Takada, (1998), “Markov network processes with product form stationary distribution”, *Queueing Systems: Theory and Applications*, 28, 377-401.
17. X. Chao and M. Pinedo, (1995), “Queueing Networks with Signals and Stage Dependent Routing”, *Probability in the Engineering and Informational Sciences*, 9, 341-354.
18. X. Chao and M. Pinedo, (1995), “On Networks of Queues with Batch Services, Signals, and Product Form Solution”, *Operations Research Letters*, 237-242.
19. X. Chao, M. Pinedo, and D. Shaw, (1996), “An Assembly Network of Queues with Product Form Solution”, *Journal of Applied Probability*, 33, 858-869.
20. X. Chao and S. Zheng, (1998), “A result on networks of queues with customer coalescence and state dependent signaling”, with S. Zheng *Journal of Applied Probability*, Vol 35, 151-164.
21. X. Chao and S. Zheng, (2000), “Triggered Concurrent Batch Arrivals and Batch Departures in Queueing Networks”, *Discrete Event Dynamic Systems,* 10, 115-129.
22. Daduna, H. (2001). “Stochastic networks with discrete time scale: Explicit expressions for the steady state behavior of discrete time stochastic networks”. Lecture Notes in Computer Science, 2046, Springer, Berlin.
23. Daduna, H. and R. Schassberger. (1983). Networks of queues in discrete time. *Zeitschrift fuer Operations Research ZOR*. 27, 159-175.
24. Gelenbe, E., (1991), “Product-form queueing networks with negative and positive customers,” *J. Appl. Prob.* 28, 656-663.
25. Glynn W.P., (1990), Diffusion Approximation, in *Handbooks on OR & MS.* D.P. Heyman and M.J. Sobel, Eds., Vol. 2, 145-198.
26. Gordon, W.J. and Newell, G. F. (1967), “Closed queueing systems with exponential servers,” *Operations Research*, 15, 254-265.
27. Harrison, J.M. and Williams R.J., (1987), “Brownian models of open queueing networks with homogeneous customer populations,” *Stochast*. 22, 77-115.
28. Harrison, J.M. and Williams R.J., (1992), “Brownian models of feedforward queueing networks: Quasireversibility and product form solutions,” *Ann. Appl. Probab.* 2, 263-293.
29. Haviv, M. (2005). “The one-chance random M/G/1 queue: A model with product form”. Working Paper. Tel-Aviv University, Israel.
30. Henderson, W., Pearce, C.E.M., Pollett, P.K. and Taylor, P.G., (1992), “Connecting internally balanced quasireversible Markov processes,” *Adv. Appl. Prob.* 24, 934-959.
31. Jackson, J.R., (1957), “Networks of waiting lines,” *Operations Research*, 5, 516-523.
32. Jackson, J.R., (1963), “Jobshop-like queueing systems,” *Management Science*, 10, 131-142.
33. Kelly, F.P., (1975), “Networks of queues with customers of different types,” *J. Appl. Prob.*, 12, 542-554,
34. Kelly, F.P., (1976), “Networks of queues,” *Adv. Appl. Prob.*, 8, 416-432.
35. Kelly, F.P., (1979), *Reversibility and Stochastic Networks,* John Wiley & Sons, New York.
36. Kelly, F.P., (1982), “Networks of quasi-reversible nodes,” in *Applied Probability-Computer Science: The Interface*, Vol.I, edited by R.L.Disney and T.J. Ott, 3-26.
37. Kingman, J.F.C., (1969), “Markov population processes,” *J. Appl. Prob.* 6, 1-18.
38. Malinkovsky, Y.V., (1990), “A criterion for pointwise independence of states of units in an open stationary Markov queueing network with one class of customers,” *Theory of Probability and Applications* 35(4) (1990), 797-802.
39. Muntz, R.R., (1972), “Poisson Departure Processes and Queueing Networks,” IBM Research Report RC4145. A shorter version appeared in Proceedings of the Seventh Annual Conference on Information Science and Systems, Princeton, 435-440.
40. Neuts, M., (1981), *Structured Stochastic Matrices of $M/G/1$ Type and Their Applications.* Marcel Dekker, New York and Basel.
41. Pollett, P.K., (1986), “Connecting reversible Markov processes,” *Adv. Appl, Prob.* 18, 880-1986.
42. Serfozo, R.F., (1989), “Poisson functionals of Markov processes and queueing networks,” *Adv. Appl. Prob.* 21, 595-611.
43. Serfozo, R.F. (1999), *Introduction to Stochastic Networks*, Springer-Verlag, New York.

44. Takada, H. and Miyazawa, M. (1997), “Necessary and Sufficient Conditions for Product-form Queueing Networks,” Science University of Tokyo.
45. van Dijk, N., (1993), *Queueing Networks and Product Forms: A Systems Approach*, Wiley & Sons, New York.
46. Walrand, J. (1988), *An Introduction to Queueing Networks*, Prentice Hall, NJ.
47. Whittle, P., (1968), “Equilibrium distributions for an open migration process,” *J. Appl. Prob.*, 5, 567-571.
48. Whittle, P., (1986), *Systems in Stochastic Equilibrium*, Wiley & Sons, New York.
49. Yashkov, S. F. (1980). “Properties of invariance of probabilistic models of adaptive scheduling in shared use systems”. *Automatic Control and Computer Science*, 14, 46-51.
50. Yates, R. D. (1994). “Analysis of discrete time queues via the reversed process”. *Queueing Systems and Their Applications*, 18, 107-116.

Chapter 6
Discrete Time Networks with Product Form Steady States

Hans Daduna

Abstract We consider networks of queues in discrete time, where the steady state distribution can be computed explicitly in closed form (product form networks): (i) Closed cycles and open tandems of single server FCFS Bernoulli nodes with state dependent service probabilities, where customers flow linearly, (ii) networks of doubly stochastic and geometrical queues (which are discrete time analogues of Kelly's symmetric, resp. general, servers), where customers of different types move through the network governed by a general routing mechanism and request for service according to general, resp. geometrical, distributions, (iii) networks with batch movements of customers and batch service, where the service and routing mechanism is defined via an abstract transition scheme.
We describe recent developments of product form networks where nodes are unreliable, break down and are repaired. This opens the possibility to investigate performance and availability of networks in an integrated model.

6.1 Introduction

Queueing network theory provided models, structural insights, problem solutions, formulas, and algorithms to many application areas. Its strong development over now around fifty years is closely connected with building a product form calculus for queueing networks in continuous time. Breakthroughs were works of Jackson [Jac57] and Gordon and Newell [GN67] in the Operations Research fields, Baskett, Chandy, Muntz, and Palacios [BCMP75] in the Computer Science, and of Kelly [Kel76]. For a survey on the state–of–the–art serve the recent books of Van Dijk [Dij93], Serfozo [Ser99], and Chao, Miyazawa, Pinedo [CMP99], previous sources are [Kel79], Whittle [Whi86], and Walrand [Wal88].

Hans Daduna
Department of Mathematics, University of Hamburg
e-mail: daduna@math.uni-hamburg.de

R.J. Boucherie and N.M. van Dijk (eds.), *Queueing Networks: A Fundamental Approach*,
International Series in Operations Research & Management Science 154,
DOI 10.1007/978-1-4419-6472-4_6, © Springer Science+Business Media, LLC 2011

Today's growing together of production, manufacturing, transportation with information processing and communication technology results in more and more complex systems which require even more elaborated models, techniques, and algorithms for better understanding their performance behaviour and for predicting performance and quality of service.

The classical single node queues and their networks are models living on a continuous time scale. And even if for some applications a discrete time scale might be more appropriate, often the well established continuous time machinery served as an approximation tool. Consequently, the survey articles [Coo90], [Wal90] from 1990 still do not review discrete time models.

But from then on an astonishing evolution of discrete time stochastic network models can be observed. Usually it is argued that the invention of ATM (Asynchronuous Transfer Mode) as the protocol for high speed transmission network technology triggered this development. Following this, special issues of *Performance Evaluation: Discrete time models and analysis methods* and *Queueing Systems and Their Applications: Advances in discrete time queues* were dedicated to the subject. The *Editorial Introductions* [TGBT94], [MT94] of these issues advertise for developing further this class of models.

Several books appeared recently dedicated to theory and applications of discrete time queueing systems and networks, [BK93], [Tak93], [Woo94], [Dad01], and in parts [CMP99].

The center of this chapter is the presentation of a discrete time analogue to the celebrated *product form calculus* of continuous time stochastic network theory. The program behind is to build a calculus which is of comparable simplicity and general applicability as the continuous time theory. Therefore this chapter refers in many parts to [Woo94] (Section 6.6.2), [CMP99] (Section 6.6.1), [Dad01] (Section 6.3 and 6.4). (To a certain extent I reused parts of [Dad01].)

I consider three classes of models:

• Linear networks (closed cycles and open tandems) of single server FCFS Bernoulli nodes.

• Networks of doubly stochastic and geometrical queues (which are discrete time analogues of Kelly's symmetric, resp. general, servers and of the BCMP nodes): Customers of different types move through the network governed by a general routing mechanism and request for service according to general, resp. geometrical, distributions.

• Networks with batch movements of customers and batch service, where the service and routing mechanism is defined via an abstract transition scheme.

I further discuss recent developments of product form networks where nodes are unreliable, break down and are repaired. This opens the possibility to investigate performance and availability of networks in an integrated model.

Parallel work on discrete time theory and several application areas are summarized in the introduction of [Dad01].

Notation: $\mathbb{R}$ denotes the real numbers, $\mathbb{R}_+ := [0,\infty)$.
The natural numbers are $\mathbb{N} := \{0,1,2,\dots\}$, the strict positive natural numbers are

$\mathbb{N}_+ := \{1,2,3,\ldots\}$, and we denote $\mathbb{Z} := \{\ldots,-2,-1,0,1,2,\ldots.\}$.
For any set A we denote by $\mathcal{P}(A)$ the set of all subsets of A.
We denote the Kronecker delta δ, resp. the complementary Kronecker delta η by

$$\delta(a,b) = \begin{cases} 1 \; if \; a = b \\ 0 \; if \; a \neq b \end{cases}, \quad \text{resp.} \quad \eta(a,b) = \begin{cases} 1 \; if \; a \neq b \\ 0 \; if \; a = b \end{cases}.$$

ACKNOWLEDGEMENT: Parts of this Chapter were written in course of a research project on STOCHASTIC NETWORKS IN DISCRETE TIME: ANALYSIS OF PERFORMANCE AND AVAILABILITY (DA774/1-1). I thank my collaborator Christian Malchin for many helpful remarks and discussions on the subject.

I thank Richard Boucherie for reading the manuscript carefully and for his help in the final version.

6.2 Bernoulli Servers with Different Customer Types and state-dependent arrivals

The Bernoulli server is the analogue of the state dependent exponential single server queue in continuous time under First–Come–First–Served (FCFS) regime. There is a single service facility where at each time instant at most one customer may be served. If at time $t \in \mathbb{N}$ a customer is in service and if there are $n-1 \geq 0$ other customers present then this service ends in the time segment $[t,t+1)$ with probability $p(n) \in (0,1)$ and the customer will depart at the end of this time slot; with probability $q(n) = 1 - p(n)$ this customer will stay at least one further time quantum. The decision for a customer whether to stay or to leave is made independently of anything else other than the queue length at time t.

All customers share the same countable type set M (“single chain” case). The type of an arriving customer is chosen as follows: Arrival probabilities depend on the history of the system only through the actual queue length, i.e., if at time t there are n customers present, then a new arrival of type m appears in $(t,t+1]$ with probability $b(n)\cdot a(m) \in (0,1)$. With probability $c(n) = 1 - b(n)$ there will be no arrival. Such an arrival stream will be termed henceforth *state dependent Bernoulli arrival process.* (Sometimes we allow $p(n) = 1$ and/or $b(n) = 1$.)

Departures and arrivals occur conditionally independent given the actual queue lenght. Joint arrivals and departures are scheduled according to LA-D/A regime (*late arrivals–departure before arrivals*) [GH92], see Figure 6.1.

If at a customer’s arrival instant the server is free her service immediately commences. Otherwise she enters the waiting room which is organized on a FCFS basis (sometimes called FIFO: First–In–First–Out). If a customer has obtained her total service request she immediately departs from the system. If a customer departs and there is at least one further customer present then the customer at the head of the waiting line enters the server, her service commences immediately, and all other waiting customers are shifted one place up in the line. The time needed for reorganizing the queue is assumed neglectible (zero time).

The system's development over time is described by a discrete time Markov chain $X = (X(t) : t \in I\!N)$. The state is recorded at times $t \in I\!N$ just after possible departures $D(t)$ and arrivals $A(t)$ have happened (Figure 6.1). A typical state of the system is described by a type sequence $x = (x_1, \dots, x_n) \in M^n$, where for $n > 0$ x_1 is the type of the customer in service, x_2 is the type of the customer at the head of the queue,..., x_n is the type of the customer who arrived most recently. The empty system is denoted by $x = e$. (We set for the empty system the queue length $n = 0$.) Let $X(t)$ denote the state of the node at time $t \in I\!N$.

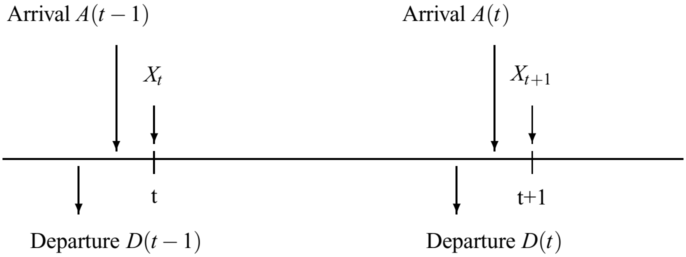


Fig. 6.1: Regulation of arrivals and departures

$X = (X(t) : t \in I\!N)$ is irreducible with state space $\tilde{S} := \{e\} \cup \bigcup_{n=0}^{\infty} M^n$.

Theorem 6.2.1 (Steady state) *If the Markov chain X is ergodic then the unique equilibrium distribution of X is with norming constant $H < \infty$*

$$\begin{aligned}\pi(x) &= \pi(x_1, \dots, x_n) \qquad\qquad x = (x_1, \dots, x_n) \in \tilde{S}.\\ &= \left(\frac{\prod_{m=0}^{n-1} b(m)}{\prod_{m=0}^{n} c(m)}\right) \left(\prod_{k=1}^{n} a(x_k)\right) \left(\frac{\prod_{m=1}^{n-1} q(m)}{\prod_{m=1}^{n} p(m)}\right) \cdot H^{-1}. \qquad (6.1)\end{aligned}$$

Remark 6.1 (Steady state decomposition). π shows a decomposition (separation) of steady states into factors concerning arrival, service, and type selection probabilities. Such separability is common to almost all product form steady states in continuous time and occurs in discrete time queueing networks as well.

Theorem 6.2.2 (Arrival Theorem) *Let X be in equilibrium and denote by*
$A(m,t) = \{$at time t a customer of type m arrives at the node$\}$
the arrival event of interest. Then for $x = (x_1, \dots, x_n) \in \tilde{S}$,

$$\begin{aligned}\pi_{1,m}(x) &:= P(X(t) = (x_1, \dots, x_n, m) | A(m,t)) \qquad (6.2)\\ &= \left(\frac{\prod_{m=0}^{n} b(m)}{\prod_{m=0}^{n+1} c(m)}\right) \left(\prod_{k=1}^{n} a(x_k)\right) \left(\prod_{m=1}^{n} \frac{q(m)}{p(m)}\right) H_1^{-1}.\end{aligned}$$

The norming constant H_1 does not depend on the arriving customer's type.

The interpretation of $\pi_{1,m}$ is that it describes the distribution of the other customers' disposition in a type-m arrival instant under equilibrium conditions. Remarkable is that this arrival distribution has not the form of the equilibrium distribution even if the arrival process is a state independent Bernoulli process. In continuous time it is true for state independent arrivals that time stationary and customer arrival stationary distribution coincide. For systems with Poisson arrivals this is the PASTA property (Poisson Arrivals See Time Averages) [Wol82].

From a general point of view the PASTA theorem and its relatives determine the stationary and asymptotic distribution of systems when the observation points are prescribed by an associated (embedded) point process, for a review see [BB94], chapter 4, section 3. Palm theory in discrete time ([BB94], Chapter 1, Section 7.4) yields similar results via elementary conditional probabilities.

In discrete time, a PASTA analogue usually does not hold, although exceptions can be found. An early result was proved by Halfin in [Hal83]. Characterisation theorems of the PASTA type (thereby strengthening the BASTA–results (Bernoulli Arrivals See Time Averages) from [MMW89]) were proved by El-Taha and Stidham [ETS92], (see also [ETS99], section 2, theorem 3.18 and corollary 3.19). Miyazawa and Takahashi [MT92] proved ASTA in a discrete time point process setting by using a rate conservation principle. They also observed that for some systems this property does not hold.

Corollary 6.2.3 (End–to–end–delay) *[Dad01][Theorem 2.12] Consider the Bernoulli server with state dependent arrival rates $b(n) \in (0,1)$ and state independent service rates $p(n) = p \in (0,1)$ in equilibrium with a test customer of type m arriving at time 0 finding the other customers distributed according to $\pi_{1,m}$, see* (6.2). *Denote by $P_{\pi_{1,m}}$ a probability measure which governs X under this conditions and by $E_{\pi_{1,m}}[\cdot]$ expectations under $P_{\pi_{1,m}}$.*
Denote by S the test customer's sojourn time in system. Then with (see (6.1)*)*

$$\alpha(\theta) = \sum_{n=0}^{\infty} \frac{\prod_{m=0}^{n-1} b(m)}{\prod_{m=0}^{n} c(m)} \theta^m, \quad |\theta| \le q/p, \tag{6.3}$$

$$E_{\pi_{1,m}} \theta^S = \left(\alpha(\frac{q\theta}{1-q\theta}) - \alpha(0)\right) \cdot \left(\alpha(\frac{q}{p}) - \alpha(0)\right)^{-1}, \quad |\theta| \le 1. \tag{6.4}$$

Corollary 6.2.4 (Queue length process) *The queue length process is a homogemeous Markov chain, which we denote by X as well. If X is ergodic, then its unique stationary and limiting distribution is (with $H < \infty$ from* (6.1)*)*

$$\pi(n) = \frac{\prod_{m=0}^{n-1} b(m)}{\prod_{m=0}^{n} c(m)} \cdot \frac{\prod_{m=1}^{n-1} q(m)}{\prod_{m=1}^{n} p(m)} \cdot H^{-1}. \quad n \in \mathbb{N}, \tag{6.5}$$

If additionally $p(n) = p \in (0,1)$, and $b(n) = b \in (0,1), n \in \mathbb{N}$, then X is ergodic if and only if $b < p$, and if this holds the stationary distribution of X is

$$\pi(n) = (1 - \frac{b}{p})\left(\frac{bq}{cp}\right)^n \left(\frac{1}{q}\right)^{\eta(0,n)}, \quad n \in \mathbb{N}. \tag{6.6}$$

Remark 6.2 (Random walks in discrete time). The queue length of the Bernoulli server is a random walk on $\mathbb{N}$ (with reflection at 0) in discrete time in the sense of [KSK76], p. 84, or a birth and death chain in discrete time, see [Hun83a],(Vol. I), p. 178. Corollary 6.2.4 and the corollaries below are therefore simple consequences of the limiting and stationary behaviour of birth and death chains, see [Hun83b], (Vol. II), Example 7.2.2, p. 107.

Remark 6.3 (Reversibility). For state independent arrival probabilities, Hsu and Burke [HB76] proved that in steady state the queue length process X is time reversible. So, in equilibrium the departure process is a Bernoulli–(b) process, and the departure process up to t and the state at t are independent. This lead Hsu and Burke to apply separability to tandem queues.

In [CMP99], example 12.10, and the remark below on p.354, it is shown that this queue is quasi–reversible according to the definition 12.6 there.

The system dealt with in theorem 6.2.1 is neither reversible nor quasi–reversible.

Corollary 6.2.5 (Loss systems) *Assume that in the setting of corollary 6.2.4 we have $b(n) \in (0,1)$ for $n \leq L-1 > 0$, and $b(n) = 0$ for $n \geq L$. Then X is ergodic on $E = \{0,1,\ldots,L\}$, and the stationary distribution of X is*

$$\pi(n) = \frac{\prod_{m=0}^{n-1} b(m)}{\prod_{m=0}^{n} c(m)} \cdot \frac{\prod_{m=1}^{n-1} q(m)}{\prod_{m=1}^{n} p(m)} \cdot H^{-1}. \quad n \in E, \tag{6.7}$$

For the discrete time $M/M/s/\infty$ no simple closed form expressions for the steady state are at hand. Usually root solving procedures for multidimensional boundary equations are applied. Related problems are dealt with in [BSDP92], [DT92] [SZ94], [BK93], section 4.1.2. For a light traffic approximation see [Dad01][example 2.10]. The no-waiting-room case is considered in [CG96]. Multiserver queues in discrete time for modeling controlled ATM switches are described in [RMW94], where a leaky–bucket control is investigated.

Pestien and Ramakrishnan [PR94b], section 3, proved that including $\cdot/M/s/\infty$ into a closed cycle of queues destroys the product form equilibrium if $1 < s < \infty$.

Feedback queues are models of repeated visits to a production or service facility, and rework of an item, e.g., with production control. ATM transmission systems are described in [STH98], where in a node with service time deterministic-(1) the feedback mechanism models successive transmission of cells of a message, with geometrically distributed length. Feedback destroyes the FCFS structure of the systems, which reflects real systems' protocol behaviour. Related is *Round–Robin* regime , described e.g. in [Kle64], [LB96] (see the references there).

We consider a feedback node where customers of different types from the set M of possible types are served, i.e., the model of section 6.2 with a queue length dependent Bernoulli feedback. The state space is $\tilde{S}$, defined before theorem 6.2.1. A customer departing from the queue leaving behind $m-1$ customers is fed back into the waiting room (to the tail of the queue) with probability $r(m)$. If she was the only customer present she will obtain immediately a further service, otherwise she will join the tail of the queue. With probability $1-r(m)$ she will leave the system. The decision whether to leave or to reenter is made independently of anything else. The regulation of customer movements in case of multiple events is: Departure before arrival for a joint arrival and departure (D/A) and feedback before arrival for a joint feedback and arrival (F/A).

Theorem 6.2.6 (Feedback queue with customer types) *If on state space $\tilde{S}$ the Markov chain is ergodic, then its steady state is for $(x_1,\dots,x_n) \in \tilde{S}$,*

$$\pi(x_1,\dots,x_n) = \frac{\prod_{m=0}^{n-1} b(m)}{\prod_{m=0}^{n} c(m)} \cdot \prod_{k=1}^{n} a(x_k) \cdot \frac{\prod_{m=1}^{n-1}(q(m)+p(m)r(m))}{\prod_{m=1}^{n} p(m)(1-r(m))} \cdot H^{-1}.$$

Setting $p(n)=1, n \geq 1, r(m)=r, m \in I\!N$ yields the round–robin scheme of [STH98], which shows some further features not represented here. In [LB96] the number of packets arriving per slot is an of i.i.d. sequence, service time is of phase type, and the service mechanism is round–robin. Product form steady state occurs in [DS81].

6.3 Closed Cycles of Bernoulli Servers

In this chapter we construct closed cycles of state dependent Bernoulli servers with a fixed number of customers cycling. We have multiple customer types as described in section 6.2. In section 6.3.1 we determine steady state behaviour and the individual customers' behaviour at arrival instants at the nodes. We concentrate on the unichain case (all customers share the same set of possible types) throughout. The multichain case is sketched in [Dad01][Section 3.2]. In section 6.3.2 we determine a travelling customer's sojourn time distributions at the nodes in a cycle. We end with explicit expressions for the generating function (z-transform) of the vector of the successive sojourn times which can be inverted easily by direct methods. We sketch in section 6.3.3 algorithms to compute different norming constants.

The first explicit result on the steady state behaviour for closed cycles of state independent Bernoulli servers appeared in 1994 – see [PR94b] and [PR94a]. The parallel result for open series of such nodes is already from 1976, see [HB76]. A shorter proof of the steady state result can be found in [Dad97c].

6.3.1 Steady State Behaviour and Arrival Theorem

Consider a closed cycle of state–dependent Bernoulli servers under FCFS queueing regime with unlimited waiting room. There are J nodes in the cycle, numbered

$1,2,\dots,J$; customers leaving node j procede immediately to node $j+1$. We formally define for the numbering of the nodes $J+1 := 1$ and $1-1 := J$.

K customers cycle in the system being of types $m \in M$, $|M| < \infty$. Customers may change their types according to a Markovian rule when departing from a node. All customers show the same behaviour with respect to their types.

With probability $r(i;m,m') \geq 0$ a customer of type m on leaving node i becomes a customer of type $m' \in M$ when entering node $i+1$, $i = 1,\dots,J$. Given i and m, the selection of m' is done independently of anything else in the past of the system. We assume that the system of equations

$$\eta(i;m) = \sum_{m' \in M} \eta(i-1;m')r(i-1;m',m), i = 1,\dots,J, m \in M, \tag{6.8}$$

has a unique stochastic solution $\eta = (\eta(i;m) : i = 1,\dots,J, m \in M)$.

The evolution of the system is described by a multivariate Markov chain $X := (X(t) : t \in \mathbb{N})$ as follows: Let $M(i) = \{m \in M : \eta(i;m) > 0\}$ denote the set of possible types which customers may show when staying in node $i, i = 1,\dots,J$. A typical state of the system is denoted by $x = (x_1,\dots,x_J)$, where $x_j = e_j$ or $x_j = (x_{j1},\dots,x_{jn_j}) \in M(j)^{n_j}, 1 \leq n_j \leq K, j = 1,\dots,J$, and $n_1 + \dots + n_J = K$. x_j is called a local state for node j with the following meaning :

If $x_j = e_j$ then node j is empty and we set $n_j = 0$.

If $x_j = (x_{j1},\dots,x_{jn_j}), n_j > 0$, then a customer of type x_{j1} is in service at node j, x_{j2} is the type of the customer waiting at the head of the queue,..., and x_{jn_j} is the type of the customer who arrived most recently at node j. The local states are concatenated to global states in the state space

$$\tilde{S}(K,J) := \{(x_1,\dots,x_J) : x_j = (x_{j1},\dots,x_{jn_j}) \in M(j)^{n_j}, 1 \leq j \leq J, \sum_{j=1}^{J} n_j = K\}$$

where $X = (X(t) := (X_1(t),\dots,X_J(t)) : t \in \mathbb{N})$ is living on.

The nodes operate independently as follows: If at time t at node j a customer is in service and if there are $n_j - 1 \geq 0$ other customers present at that node then this service ends in the time segment $[t,t+1)$ with probability $p_j(n_j) \in (0,1)$ and the departed customer will be at the end of the queue of node $j+1$ at time $t+1$; with probability $q_j(n_j) = 1 - p_j(n_j)$ this customer will stay at least one further time quantum at node j, $j = 1,\dots,J$. Whether a customer stays on or to leaves node j is independent of the history given the local state of X at j. A customer arriving at node $j+1$ at time $t+1$ either joins the end of the queue there (if other customers are present) or immediately enters service (if at time t node j was empty or there has been exactly one customer who obtained her last quantum of service time). If at some node at the same epoch an arrival and a departure occur we always assume that the departure event takes place first, D/A–rule; see Figure 6.1 and [GH92].)

The state of the system is recorded at times $t \in \mathbb{N}$ just after possible departures and arrivals had happened. (Due to ample waiting room the *A/D–rule*, arrival before departure, yields the same steady state distribution.) The states which X will enter may be a subset of $\tilde{S}(K,J)$, depending on the initial state of the process. We set the following assumption in force.

Assumption 6.3.1 (Irreducibility) *Depending on the initial state X is irreducible on its state space. We denote the state space in any case by $\tilde{S}(K,J)$, the meaning of which will be clear from the context.*

Theorem 6.3.2 (Steady–State Distribution) $X = (X(t) : t \in I\!N)$ *is positive recurrent and its unique steady state is with norming constant* $G(K,J)^{-1}$

$$\pi^{K,J}(x_{11},\dots,x_{1n_1};\dots;x_{J1},\dots,x_{Jn_J}) \qquad (x_{11},\dots,x_{Jn_J}) \in \tilde{S}(K,J)$$
$$= \prod_{j=1}^{J}\left(\prod_{k=1}^{n_j}\eta(j;x_{jk})\right)\left(\frac{\prod_{h=1}^{n_j-1} q_j(h)}{\prod_{h=1}^{n_j} p_j(h)}\right)\cdot G(K,J)^{-1}. \tag{6.9}$$

For obvious reasons the steady state distributions $\pi^{K,J}$ are said to be of *product form*, which clearly does not imply independence of the local queue lengths. n interesting difference between (6.9) and the continuous time analogue (see e.g. [Kel79]), is clarified when considering state independent service rates.

Corollary 6.3.3 *Suppose we have $p_j(n_j) = p_j, n \in I\!N_+, j = 1,\dots,J$. Then the unique stationary distribution of X is for* $(x_{11},\dots,x_{Jn_J}) \in \tilde{S}(K,J)$,

$$\pi^{K,J}(x_{11},\dots,x_{1n_1};\dots;x_{J1},\dots,x_{Jn_J}) \tag{6.10}$$
$$= \prod_{j=1}^{J}\left(\prod_{k=1}^{n_j}\eta(j;x_{jk})\right)\left(\frac{1}{q_j}\right)^{\eta(0,n_j)}\left(\frac{q_j}{p_j}\right)^{n_j}\cdot G(K,J)^{-1}.$$

The extra factor $\left(\frac{1}{q_j}\right)^{\eta(0,n_j)}$ for non empty queues sets the difference. Therefore in the homogeneous cycle ($p_j = p, j = 1,\dots,J$) the stationary distribution is not uniform on $\tilde{S}(K,J)$ as in continuous time. For consequences see [PR94a].

Corollary 6.3.3 and even the case of state dependent service probabilities can in principle be proved by applying theorem 2.2 of [HT91] or by similar derivations as presented in section 4 of [BD91].

Remark 6.4 (Bottleneck behaviour). Consider the closed cycle of corollary 6.3.3 with state independent service probabilities and $p_1 < \min(p_2,\dots,p_M)$, i.e. node 1 is the unique slowest server and will act as bottleneck of the network. This means: If we have a cycle with large population size ($K \gg J$) then in equilibrium (and asymptotically over time) we shall see with high probability almost all customers at node 1 and node 1 acts asymptotically as a Bernoulli source.

More precisely: For fixed state $(x_{21},\dots,x_{2n_2};\dots;x_{J1},\dots,x_{Jn_J}) \in \times_{j=2}^{J} M(j)$ at nodes $2,\dots,J$ denote by $\pi^{K,J}(x_{21},\dots,x_{2n_2};\dots;x_{J1},\dots,x_{Jn_J})$ the marginal probability for the coordinates $2,\dots,J$ under $\pi^{K,J}(\cdot)$ in a cycle with J nodes and $K \geq n_2 + \dots n_J$ customers. Then we have

$$\lim_{K\to\infty} \pi^{K,J}(x_{21},\dots,x_{2n_2};\dots;x_{J1},\dots,x_{Jn_J}) = \tag{6.11}$$
$$= \prod_{j=2}^{J}\left(\prod_{k=1}^{n_j}\eta(j;x_{jk})\right)\left(\frac{1}{q_j}\right)^{\eta(0,n_j)}\left(\frac{p_1 q_j}{q_1 p_j}\right)^{n_j}\cdot\left(1-\frac{p_1}{p_j}\right).$$

The limiting probability is the stationary distribution of an open tandem of nodes $2,\dots,J$ with a Bernoulli arrival stream with parameter p_1, see theorem 6.4.1.

Corollary 6.3.4 *[DPR03] For each node j, we define the throughput $T^{K,J}_{(j)}$ at any node j as the expected progress at node j at a given instant under $\pi^{K,J}$:*

$$T^{K,J}_{(j)} = \sum_{(x_1,\dots,x_J)\in\tilde{S}(K,J)} \pi^{K,J}(x_1,\dots,x_J)\cdot p_j(n_j). \tag{6.12}$$

$T^{K,J} := T^{K,J}_{(j)}$ is independent of j.
If the service probabilities $p_j(m)$ are nondecreasing in $m \in \mathbb{N}$ for all j, $T^{K,J}$ is nondecreasing in K and nonincreasing in J.

In continuous time a consequence of the product form is the *Arrival Theorem*, [LR80], [SM81], which roughly states, that in equilibrium an arriving customer at node j observes the other customers distributed according to the equilibrium of the system if she himself would not be there, $j = 1,...,J$. So the *arrival distribution* has the same structure as the equilibrium, and is independent of the node j, where the customer arrives. This is not the case in discrete time.

Property 6.3.5 (Arrival Theorem) *[Dad96] Let $X = (X(t) : t \in \mathbb{Z})$ be the stationary continuation of $X = (X(t) : t \in \mathbb{N})$ under $\pi^{K,J}$. Assume that for node i and customer type m there exists some m' such that $r(i-1;m',m) > 0$, i.e. $m \in M(i)$, and denote by $A(i,m)$ the event that at time 0 an arrival of a type–m customer at node i appeared, $i \in \{1,\dots,J\}$. Then for $x = (x_1,\dots,x_J) \in \tilde{S}(K-1,J)$ with $G_{i,m}(K,J)^{-1}$ as norming constant*

$$\begin{aligned}&\pi^{K,J}_{i,m}(x_1,\dots,x_J) \qquad\qquad (6.13)\\ &:= P(X(0) = (x_1,\dots,x_{i-1},(x_{i,1},\dots,x_{i,n_i},m),x_{i+1},\dots,x_J)|A(i,m))\\ &= \left(\prod_{k=1}^{n_i}\eta(i;x_{ik})\right)\left(\frac{\prod_{h=1}^{n_i} q_i(h)}{\prod_{h=1}^{n_i} p_i(h)}\right)\\ &\quad\cdot\prod_{j=1,j\neq i}^{J}\left(\prod_{k=1}^{n_j}\eta(j;x_{jk})\right)\left(\frac{\prod_{h=1}^{n_j-1} q_j(h)}{\prod_{h=1}^{n_j} p_j(h)}\right)\cdot G_{i,m}(K,J)^{-1}.\end{aligned}$$

A different form of an arrival theorem was proved by Henderson and Taylor [HT91]: They computed in a general setting the disposition probability for stay–on customers seen by a prescribed set of departing customers just before the latter enter their destination node. These probabilities can be computed similarly to the proof of proposition 6.3.5. And the other way round: Having the arrival probabilities of

[HT91], Corollary 2.4, at hand, by a suitable, but lengthy, summation our result for indistinguishable customers would follow.

6.3.2 Delay Times for Customers in a Closed Cycle

For a closed cycle of J state independent Bernoulli servers as described in section 6.3.1 and specified in corollary 6.3.3 we derive the steady state cycle time for a customer in the system, and similar quantities. Most important is to determine the joint distribution of successive sojourn times (waiting time + service time) of a customer at the different nodes during such a cycle.

From the arrival theorem 6.3.5 and the type independent service times it follows that these distributions do not depend on the type of the cycling customer. We therefore can and will restrict our attention to the case of indistinguishable customers. The joint queue length process $X = (X(t) : t \in \mathbb{N})$ is Markov with state space $S(K,J) = \{(x_1,\dots,x_J) \in \mathbb{N}^J : x_1 + \dots + x_J = K\}$.

$X(t) = (X_1(t),\dots,X_J(t)) = (x_1,\dots,x_J)$ indicates that at time t there are x_j customers present at node j, including the one in service, if any, $j = 1,\dots,J$.

Consider a test customer C_0 arriving at time 0 at node i finding the other customers distributed according to $\pi_i^{K,J}$. Denote by $P_{\pi_i^{K,J}}$ a probability that governs the system with this condition, and by $E_{\pi_i^{K,J}}[\cdot]$ expectations under $P_{\pi_i^{K,J}}$.

Theorem 6.3.6 (Joint sojourn time distribution) *[Dad97b] If C_0 arrives at time 0 at node $i \in \{1,\dots,J\}$, finding the other customers distributed according to $\pi_i^{K,J}$, and if $(S_1^{(i)},S_2^{(i)},\dots,S_J^{(i)})$ denotes the vector of her sojourn times during her cycle which starts at 0, then*

$$E_{\pi_i^{K,J}}\left[\prod_{j=1}^{J}\theta_j^{S_j^{(i)}}\right] \qquad |\,\theta_j\,| \leq 1, j = 1,\dots,J \tag{6.14}$$

$$= \sum_{(x_1,\dots,x_j)\in S(K-1,J)} G_i(K,J)^{-1}\left\{\prod_{j=1}^{J}\left(\frac{p_j\theta_j}{1-q_j\theta_j}\right)\right\}$$

$$\cdot\left(\frac{q_i\theta_i}{1-q_i\theta_i}\right)^{x_i}\prod_{j=1,j\neq i}^{J}\left\{\left(\frac{1}{q_j\theta_j}\right)^{\eta(0,x_j)}\left(\frac{q_j\theta_j}{1-q_j\theta_j}\right)^{x_j}\right\},$$

The joint distribution of C_0's successive sojourn times vector $(S_1^{(i)},\dots,S_J^{(i)})$ during this cycle does **not** *depend on the node i, where the cycle started.*

The appealing interpretation of the RHS of (6.14) as a direct result of conditioning on the arrival situation is false, with $i = 1$ we note:

$$\left(\frac{p_1\theta_1}{1-q_1\theta_1}\right)^{x_1+1}\prod_{j=2}^{J}\left\{\left(\frac{p_j\theta_j}{1-q_j\theta_j}\right)^{x_j+1}\left(\frac{1}{\theta_j}\right)^{\eta(0,x_j)}\right\}$$

is **not** the conditional distribution $\mathcal{L}(S_1^{(1)},\ldots,S_J^{(1)} \mid X(0) = (x_1+1,x_2,\ldots,x_J))$.

Note that in (6.14) the asymmetry of (6.13) (which is substantial there) formally reappears, but nevertheless we have symmetry in $i \in \{1,\ldots,J\}$.

Remark 6.5. The proof of theorem 6.3.6 is performed by induction on the length of the cycle and the number of customers. We consider the cycle built of an initial node connected to a smaller residual cycle with less customers cycling, where the induction hypothesis applies. Insofar the proof is standard, mimicing the continuous time analogue [KP83]. Kelly and Pollett showed even more: Splitting of the cycle for the induction step is possible between every pair of nodes j and $j+1$, $j \in \{1,2,\ldots,J-1\}$ - the induction step is always the same ! The main difficulty that arises in the discrete time system is to justify this splitting. Directly carrying over the arguments from [KP83] is not possible, because of the dependence of the disposition distribution for the other customers on the node where the customer jumps.

Remark 6.6 (End–to–end–delay). As a by-product of theorem 6.3.6 we have a result on the distribution of cycle times (= *sum of successive sojourn times*) which is the end–to–end–delay e.g. in a transmission line under window flow control for a system in heavy traffic (see [Rei79] and [Rei82]): Put $\theta_j = \theta, j = 1,\ldots,J$. A direct proof is given in [Dad96].

Theorem 6.3.6 allows to compute moments and covariances explicitly. Results on partial-cycle times are given in [Dad96], section 4.

Remark 6.7 (An invariance property versus bottleneck behaviour). The discussion of the bottleneck behaviour under $p_1 < \min(p_2,\ldots,p_M)$, i.e. when node 1 is the bottleneck of the network, in remark 6.4 suggests that for $(K \gg J)$ the cycle time will be determined almost completely by the sojourn time of the test customer at node 1. This is made precise by Boxma [Box88] in continuous time. Boxma's result in the discrete time setting is for $K \longrightarrow \infty$

$$E_{\pi_1^{K,J}}[S_1^{(1)} + \ldots + S_J^{(1)}] = Kp_1^{-1}\left\{1 + \mathcal{O}\left(\frac{p_1}{\min(p_2,\ldots,p_M)}\right)^K\right\}, \tag{6.15}$$

(even $o(\cdot)$ instead of $\mathcal{O}(\cdot)$ can be shown) and it is also $\lim_{K\to\infty} K^{-1}E_{\pi_1^{K,J}}[S_1^{(1)}] = p_1^{-1}$. So the overwhelming part of the customer's cycle time is her visit at the bottleneck. Therefore the following observation is striking ([MD05] for continuous time): If we compute the conditional joint counting densities of the successive sojourn times of a cycling customer given her cycle time, then for all feasible values this conditional densities are independent of K [MD04].

6.3.3 Computational Algorithms for Closed Cycles of State Independent Bernoulli Servers

Algorithms are developed for continuous time networks to efficiently evaluate norming constants, steady state probabilities, and further quantities derived from this. An analysis of convolution algorithm (norming constants), Mean Value Analysis (MVA), the more recent RECAL method, and further approximation algorithms for product form networks as well as for non product form networks based on those concepts is in [BGdT98]. For a short survey, including approximation procedures, see [HNS99]. Algorithms for discrete time systems with finite capacity, are developed in [SG97].

We sketch in this section the prolems occurring due to the discrete time scale. A first observation is based on theorem 6.3.2 and proposition 6.3.5: While in continuous time models the norming constants in the steady state and in the arrival probabilities are of the same structure, we have to apply different computational schemes. We restrict ourself to state-independent service rates.

Property 6.3.7 (Norming constants) *(a) For the steady state distribution (6.9) in theorem 6.3.2 (with state independent service probabilities) the norming constant in a system with K customers cycling in J nodes is*

$$G(K,J) = \sum_{\substack{(n_1,\dots,n_J)\in\mathbb{N}^J\\ n_1+\dots+n_J=K}} \prod_{j=1}^{J} \left(\frac{q_j}{p_j}\right)^{n_j} \left(\frac{1}{q_j}\right)^{\eta(0,n_j)}, \quad K \geq 1, J \geq 1.$$

(b) The norming constant for the arrival probabilities (6.13) in proposition 6.3.5 seen by a type m–customer on his arrival at node i in a system with K customers cycling in J nodes is

$$G_{i,m}(K,J) = \sum_{\substack{(n_1,\dots,n_J)\in\mathbb{N}^J\\ n_1+\dots+n_J=K-1}} \left(\frac{q_i}{p_i}\right)^{n_i} \prod_{\substack{j=1\\ j\neq i}}^{J} \left(\frac{q_j}{p_j}\right)^{n_j} \left(\frac{1}{q_j}\right)^{\eta(0,n_j)}, \quad K \geq 1, J \geq 1.$$

The constants are type independent. We set $G_i(K,J) := G_{i,m}(K,J)$, $m \in M$.

The following is an analogue of *Buzen's algorithm* ([Buz73]).

Property 6.3.8 (Buzen's algorithm) *For $K \geq 1, J \geq 1$ and $p_j \in (0,1), q_j = 1 - p_j, j = 1,\dots,J$, the following recursion holds for $G(K,J)$:*

$$G(1,J) = \sum_{j=1}^{J} \frac{1}{p_j}, \qquad J \geq 1, \qquad\qquad G(K,1) = \left(\frac{q_1}{p_1}\right)^K \frac{1}{q_1}, \qquad K \geq 1,$$

$$G(K,J) = G(K,J-1) + \frac{q_J}{p_J} G(K-1,J) + G(K-1,J-1), \quad K \geq 2, J \geq 2. \tag{6.16}$$

Note that the computation of (6.16) depends on the prescribed numbering of the nodes. Renumbering the nodes leads to different paths for the algorithm.

The norming constant for the arrival probabilities will be computed for any customer arriving at node 1. It does not depend on the type of the arrival. Furthermore: From [Dad01], lemma 7.2, it follows that for all $j = 1,\dots,J$ equality $G_i(K,J) = G_1(K,J)$ holds.

Computing norming constants and performance indices needs an interplay of constants of different type: from time stationary steady states and from customer stationary steady states. Such distinction is not necessary in continuous time.

Property 6.3.9 *For $K \geq 1, J \geq 1$ and $p_j \in (0,1), q_j = 1 - p_j, j = 1,\dots,J$, the following recursion holds for $G_1(K,J)$:*

$$G_1(2,J) = \frac{q_1}{p_1} + \sum_{j=2}^{J} \frac{1}{p_j}, \quad J \geq 1,$$

$$G_1(K,J) = G(K-1,J-1) + \frac{q_J}{p_J} G_1(K-1,J), \qquad K \geq 3, J \geq 1.$$

Some elementary consequences of the above algorithms follow.

Corollary 6.3.10 *For a random vector $(X_1,\dots,X_J)$ distributed according to the equilibrium distribution $\pi^{K,J}$ in the closed cycle (see corollary 6.3.3) holds:*
(a) The probability for queue length X_1 at node 1 to exceed $k \in \{0,1,\dots,K\}$ is

$$P(X_1 \geq k) = \left(\frac{q_1}{p_1}\right)^k \frac{1}{q_1} G_1(K-k,J) G(K,J)^{-1}$$

(b) The mean steady state queue length $E[X_1]$ at node 1 is

$$E[X_1] = \frac{1}{q_1} G(K,J)^{-1} \sum_{k=1}^{K} \left(\frac{q_1}{p_1}\right)^k G_1(K-k,J).$$

(c) For $j = 1,\dots,J$ and $k \geq 0$ we have

$$P(X_j \geq k) = P(X_1 \geq k) \left(\frac{p_1 q_j}{q_1 p_j}\right)^k \frac{q_1}{q_j}$$

6.3.4 Large Cycles of State Dependent Bernoulli Servers

In the remarks 6.4 and 6.7 the number of nodes J in the cycles was fixed while the number of customers grew unboundedly. The limiting behaviour of the sequence of networks provides information about the behaviour of the cycle when the system is overloaded by high population sizes. A related question about approximating the behaviour of large networks can be studied by observing sequences of networks where the number of nodes and the number of customers grow simultaneously. Pestien and Ramakrishnan studied the asymptotic throughput and queue lenghts in networks (indexed by N) of state independent nodes, where the number of nodes fulfill $J(N) \longrightarrow \infty$ and the limit of the customer/nodes ratios

$$\alpha = \lim_{N\to\infty} K(N)/J(N) \in [0,\infty]$$

exists, [PR94a], [PR99], In [DPR03] this is extended to the model in theorem 6.3.2 (with undistinguishable customer).

We have L types of nodes, characterized by distinct nondecreasing sequences of success probabilities $p_{[0]},\dots,p_{[L-1]}$, i.e., for each ℓ, $p_{[\ell]}(h)$ is the service probability at a node of type ℓ when the queue length is h. Let the maximal service rate for type be

$$p^*_{[\ell]} = \lim_{h\to\infty} p_{[\ell]}(h),$$

and assume $0 < p^*_{[L-1]} \le p^*_{[L-2]} \le \cdots \le p^*_{[0]} \le 1$.

Even though we assume that the service rates $p_{[\ell]}$ are distinct, we allow the possibility that $p^*_{[\ell]}$ be constant in ℓ.

For each positive integer N, for each ℓ $(0 \le \ell \le L-1)$, assume that there are $J_\ell(N) \ge 1$ nodes of type ℓ. Also suppose that for each ℓ, the proportion of nodes of type ℓ (as N approaches ∞) has a limit, which defines a density $(\beta_\ell : \ell = 0,1,\dots,L-1)$,

$$\beta_\ell = \lim_{N\to\infty} J_\ell(N)/J(N).$$

Theorem 6.3.11 *[DPR03] Let* $T(N) := T^{K(N),J(N)}$ *denote the throughput for the* N^{th} *network as defined in (6.12).*

Denote by $m_{p_{[\ell]},b}$ *the expected queue length of a stationary single node with steady state distribution* (6.5)*, where* $b(m) = b, m \in \mathbb{N}$*, is a constant arrival rate and* $p_{[\ell]} = (p_{[\ell]}(m) : m \in \mathbb{N})$ *is a state dependent service rate.*

Let g be the function defined by $g(\theta) = \sum_{\ell=0}^{L-1} \beta_\ell \cdot m_{p_{[\ell]},\theta}$*, for* θ *such that* $0 \le \theta < p^*_{[L-1]}$. *g is continuous and strictly increasing and* $\lim_{\theta\to 0} g(\theta) = 0$.

The limiting throughput, $\lim_{N\to\infty} T(N)$, *exists, and the following cases summarize the possible values of this limit:*

(i) If $\alpha = 0$, *then* $\lim_{N\to\infty} T(N) = 0$.
(ii) If $0 \le \alpha < \lim_{\theta \uparrow p^*_{[L-1]}} g(\theta)$, *then* $\lim_{N\to\infty} T(N) = g^{-1}(\alpha)$.
(iii) If $\alpha \ge \lim_{\theta \uparrow p^*_{[L-1]}} g(\theta)$, *then* $\lim_{N\to\infty} T(N) = p^*_{[L-1]}$.

For node type ℓ in network N denote by $X_{(\ell)}(N)$ a random variable distributed like the stationary queue length of a type ℓ node in this network.

Theorem 6.3.12 *Denote by* $\theta^* := \lim_{N\to\infty} T(N)$ *the limiting throughput. For* ℓ *such that* $0 \le \ell \le L-1$, *the distribution of* $X_\ell(N)$ *has the following properties:*
(i) *For every* ℓ *such that* $\theta^* < p^*_{[\ell]}$, *the distribution of* $X_\ell(N)$ *converges in total variation norm to* π *from* (6.5) *with constant arrival rate* $b(m) = \theta^*, m \in \mathbb{N}$, *and*

state dependent service rate $p_{[\ell]} = (p_{[\ell]}(m) : m \in \mathbb{N})$*. Also, for each positive integer* r*, we have convergence of the* r*th moments.*
(ii) *For every* ℓ *such that* $\theta^* = p^*_{[\ell]}$,

$$\lim_{N\to\infty} P[X_\ell(N) \geq \Delta] = 1 \quad \textit{for every } \Delta > 0.$$

Moreover the expectations diverge to ∞.

In [PR02] Pestien and Ramakrishnan proved monotonicity properties of performance measures in the cyclic queue under steady state conditions when the service rates are state-independent. These results are utilized to refine some of their previous results in [PR94a], [PR99].

6.4 Open Tandems of Bernoulli Servers with State Dependent Arrivals

In this section we investigate a series of linearly ordered nodes, fed by a state dependent arrival stream. Tandem networks are models e.g., for production lines, transmission lines in a telecommunication network, etc. Results on steady state behaviour for tandem systems (with state independent Bernoulli input and indistinguishable customers) date back to [HB76]. Hsu and Burke solved the steady state problem by proving time reversibility of a nodes' local state process in equilibrium and then using induction. A similar procedure is not possible in case of state dependent arrival processes.

We describe the steady state behaviour of state dependent tandems with different customer types in section 6.4.1. In section 6.4.2 we compute end–to–end–delay distribution for a customer traversing the tandem and the distribution of this customer's joint sojourn times at the successive nodes of his passage.

Although series systems seem to be a rather narrow class of networks, the results usually are considered to be of value for networks with more general topology as well. The technique to reduce many network problems to problems which can be solved in linear systems is called the *Method of Adjusted Transfer Rates* and is described for discrete time systems in [Dad97a].

6.4.1 Steady State and Arrival Theorem

Bernoulli servers from section 6.2 are building blocks of an open tandem of J queues. Customers of different types arrive in a state dependent Bernoulli process at node 1, and proceed through the sequence of nodes possibly changing their types, and after leaving node J, they depart from the system. All customers share the same countable type set M. If a customer of type $m \in M$ leaves node j, then this customer's type is resampled according to probability matrix $r(j)$, his new type is $m' \in M$ with probability $r(j;m,m')$, $j = 1,\dots,J$.

Regulation of simultaneous events is according to *late arrivals* and *departure before arrivals*, see Figure 6.1 and [GH92].

External arrival probabilities depend on the total population of the system and on the type of the arrival, i.e., if at time $t \in \mathbb{N}$ there are n_j customers present at node j, $j = 1,\dots,J$, then a new arrival of type m appears in $(t,t+1]$ with probability $b(n_1+\cdots+n_J)\cdot a(m) \in (0,1)$.

Service times and arrivals are conditionally independent given the actual vector of customer types at the nodes.

We use the definitions of section 6.3.1, page 276: $M(i)$ denotes the set of possible types which customers may show when staying in node $i, i = 1,\dots,J$. Here $M(1) = \{m \in M : a(m) > 0\}$, while $M(i), i = 2,\dots,J$, is determined by solving the equation (6.8) for $\eta(i;\cdot), i = 2,\dots,J$, in the present context and then setting $M(i) = \{m \in M : \eta(i;m) > 0\}$.

A typical state of the system is $x = (x_1,\dots,x_J)$, where $x_j = e_j$ or $x_j = (x_{j1},\dots,x_{jn_j}) \in M(j)^{n_j}, 1 \le n_j, j = 1,\dots,J$. x_j is called a local state for node j with the interpretation given in section 6.3, page 276. These local states allow to construct the state space for X

$$\tilde{S}(J) := \{(x_1,\dots,x_J) : x_j = (x_{j1},\dots,x_{jn_j}) \in M(j)^{n_j}, n_j \ge 0, j = 1,\dots,J\}$$

Let $X_j(t)$ denote the local state at node j, and $X(t) = (X_1(t),\dots,X_J(t))$ the joint *vector of type sequences* at time t. $X = (X(t) : t \in \mathbb{N})$ is a discrete time irreducible Markov chain with state space $\tilde{S}(J)$.

Theorem 6.4.1 (Steady state) *[Dad97c] If X is ergodic, then the unique equilibrium distribution of X is with norming constant $H(J) < \infty$*

$$\pi^J(x_1,\dots,x_J) = \pi^J((x_{11},\dots,x_{1n_1});\dots;(x_{J1},\dots,x_{Jn_J})) = \tag{6.17}$$

$$= \left(\frac{\prod_{h=0}^{n_1+\cdots+n_J-1} b(h)}{\prod_{h=0}^{n_1+\cdots+n_J} c(h)}\right)\prod_{j=1}^{J}\left(\prod_{k=1}^{n_j}\eta(j;x_{jk})\right)\left(\frac{\prod_{h=1}^{n_j-1} q_j(h)}{\prod_{h=1}^{n_j} p_j(h)}\right)\cdot H(J)^{-1},$$

$$(x_1,\dots,x_J) \in \tilde{S}(J), where\ \ c(h) := 1-b(h), h \in \mathbb{N}$$

Theorem 6.4.1 carries over to the case $b(n) \in [0,1)$ for some $n \in \mathbb{N}$, which encompass then loss systems as well, see [Dad97c], section 3.

Example 6.4.2 (Control of Bernoulli arrival) *Consider a Bernoulli arrival stream with constant intensity $B \in (0,1]$, which feeds an open tandem of Bernoulli servers. There is a Bernoulli switch at the entrance point of the network (before node 1): If the total population size of the network is n then an arriving customer is admitted with probability $\beta(n) \in (0,1]$ and is rejected and lost with probability $1-\beta(n)$. This system fits into the class of models of theorem 6.4.1 with $b(n) = B\cdot\beta(n)$. If the arrival process is sufficiently thin ($1-\beta(n)$ sufficiently high) we have ergodicity. $(\beta(n), n \in \mathbb{N})$ then is a stability control function.*

To describe a customer's delay behaviour we again need an arrival theorem.

Theorem 6.4.3 (Arrival Theorem) *For the state process X of the open tandem in equilibrium consider the event*
$A(1,m) = \{$ *at time 0 a customer of type m arrives at node 1*$\}$. *Then*

$$\begin{aligned}\pi^J_{1,m}(x_1,\ldots,x_J) &:= P(X(0) = ((x_1,m),x_2,\ldots,x_J|A(1,m)) \qquad (6.18)\\
&= P(X(0) = ((x_{11},\ldots,x_{1n_1}),m);(x_{21},\ldots,x_{2n_2});\ldots;(x_{J1},\ldots,x_{Jn_J})|A(1,m))\\
&= \left(\frac{\prod_{h=0}^{n_1+\cdots+n_J} b(h)}{\prod_{h=0}^{n_1+\cdots+n_J+1} c(h)}\right)\prod_{j=1}^{J}\left(\prod_{k=1}^{n_j}\eta(j;x_{jk})\right)\\
&\quad\cdot\left(\prod_{h=1}^{n_1}\frac{q_1(h)}{p_1(h)}\right)\prod_{j=2}^{J}\left(\frac{\prod_{h=1}^{n_j-1} q_j(h)}{\prod_{h=1}^{n_j} p_j(h)}\right)\cdot H_1(J)^{-1}, \quad (x_1,\ldots,x_J)\in\tilde{S}(J),\end{aligned}$$

$H_1(J)$ is the norming constant, which does not depend on the type of the arriving customer. For $i \neq 1$ similar formulas apply.

The usual interpretation of $\pi^J_{i,m}$ is that it describes the distribution of the other customers' disposition in an arrival instant at node i under equilibrium conditions. Remarkable is that this distribution has not the form of the equilibrium.

Corollary 6.4.4 (Individual loss probabilities) *(a) Control of Bernoulli arrival streams (example 6.4.2): Assume that a Bernoulli arrival process with arrival probability $B \in (0,1]$ is controlled by a Bernoulli switch with admission probabilities $\beta(n), n \in I\!N$. Then the loss probability for an arriving customer of type m due to rejection is*

$$p_{l,m}(J) = 1 - \frac{1}{B\cdot H(J)}\sum_{K=0}^{\infty}\prod_{h=0}^{K}\frac{b(h)}{c(h)}G_1(K,J),$$

where $G(_1K,J)$ is the norming constant for the arrival distribution at node 1 for a customer in a closed cycle of J nodes (see section 6.3.3) with K indistinguishable customers cycling. (See [Dad97c]; theorem 1.)

(b) Open loss system: If the control of the Bernoulli-(B) process is of the form $\beta(n) = 1,\ n < L$, and $\beta(n) = 0,\ n \geq L$. Then the loss probability for an arriving customer of type m due to overflow is $p_{L,m}(J) = G_1(L,J)/H(J)$.

The results are similar to computing loss probabilities at a single station with finite buffer, single deterministic-(1)-server, and Markovian arrivals [IT99].

Theorem 6.4.5 (Throughput of the tandem) *In equilibrium the throughput of the tandem is $Th(J) = H_J(J)\cdot H(J)^{-1}$, the throughput of type m customers is $a(m)\cdot Th(J)$.*

This result seems curious: Inspection of the $H_j(J)$ in theorem 6.4.3 leads to the conjecture that the value of the throughput depends on the node where it is evaluated, because the $H_j(J)$ would appear. But it can be shown, that $H_j(J)$ is independent of j, [Dad01], lemma 7.2.

Example 6.4.6 (Re–entrant lines) *Re–entrant lines are models for complex manufacturing systems with items (parts) flowing through the line which possibly request for different kind of (repeated) service and different amount of work. For modeling re–entrant lines via queueing networks see [Kum93], and the references there. If the actions on all stages are synchronized, then the production line is suitably modeled by a discrete time network of queues, for more details see examples 12.1 and 12.37 of [CMP99]. A fundamental building block of re–entrant lines is the feedback queue described in theorem 6.2.6. Building open tandems and closed cycles of such queues with different customer types is possible with obtaining explicit product form steady states, see section 4.4 in [Dad01].*

6.4.2 Delay Times for Customers in an Open Tandem

We consider a test customer of type m arriving at time 0 at node 1 who finds the other customers distributed according to $\pi^J_{1,m}$. We denote by $P_{\pi^J_{1,m}}$ a probability which governs the system with this initial condition, and by $E_{\pi^J_{1,m}}[\cdot]$ expectations under $P_{\pi^J_{1,m}}$. The following theorem states that the joint distribution of the successive sojourn times of a customer in equilibrium is distributed like a mixture of multivariate distributions with independent negative binomial marginals.

Theorem 6.4.7 (Joint sojourn time distribution in a tandem) *[Dad97c] Let $(S_1, S_2, \ldots, S_J)$ denote the vector of the test customer's successive sojourn times (= waiting time + service time) at the nodes during her passage through the tandem. The joint distribution of $(S_1, S_2, \ldots, S_J)$ is given by*

$$
\begin{aligned}
& E_{\pi^J_{1,m}}\left[\prod_{j=1}^{J} \theta_j^{S_j}\right] \qquad\qquad | \, \theta_j \, | \leq 1, j = 1, 2, \ldots, J, \qquad (6.19)\\
= & \sum_{(n_1,\ldots,n_J)\in \mathbb{N}^J} \cdot \left(\frac{\prod_{h=0}^{n_1+\cdots+n_J} b(h)}{\prod_{h=0}^{n_1+\cdots+n_J+1} c(h)} \right) \left(\frac{q_1}{p_1}\right)^{n_1} \prod_{j=2}^{J} \left(\frac{1}{q_j}\right)^{\eta(0,n_j)} \left(\frac{q_j}{p_j}\right)^{n_j} \\
& \cdot \left(\frac{p_1\theta_1}{1-q_1\theta_1}\right)^{n_1+1} \prod_{j=2}^{J} \left\{ \left(\frac{p_j\theta_j}{1-q_j\theta_j}\right)^{n_j+1} \left(\frac{1}{\theta_j}\right)^{\eta(0,n_j)} \right\} \cdot H_1(J)^{-1}.
\end{aligned}
$$

Warning: The following tempting conjecture is false

$$E_{\pi^J_{1,m}}\left[\prod_{j=1}^{J}\theta_j^{S_j}|X(0)=((x_1,m),x_2,\dots,x_J)\right]$$
$$=\left(\frac{p_1\theta_1}{1-q_1\theta_1}\right)^{n_1+1}\prod_{j=2}^{J}\left\{\left(\frac{p_j\theta_j}{1-q_j\theta_j}\right)^{n_j+1}\left(\frac{1}{\theta_j}\right)^{\eta(0,n_j)}\right\}.$$

For state independent arrivals (6.19) boils down to simple expressions.

Corollary 6.4.8 (Independent sojourn times) *[Dad97b] For Bernoulli arrivals with constant rate b the generating function of* $(S_1,S_2,\dots,S_J)$ *is*

$$E_{\pi^J_{1,m}}\left[\prod_{j=1}^{J}\theta_j^{S_j}\right]=\prod_{j=1}^{J}\frac{\left(\frac{p_j-b}{c}\right)\theta_j}{1-\left(1-\frac{p_j-b}{c}\right)\theta_j},\qquad |\,\theta_j\,|\leq 1, j=1,2,\dots,J. \tag{6.20}$$

Sojourn times are independent, geometrically distributed with parameter $\frac{p_j-b}{c}$.

(6.20) is the analogue of Burke's and Reich's results on the independence of a customer's sojourn times in an open tandem (for a review see [BD90b]).

An adhoc approximation procedure to solve complex discrete time network problems is described by Bruneel and Kim [BK93], chapter 4.1.6: The end–to–end delay of a cell on its transmission through an ATM network is computed by *assuming that the successive single node delays behave statistically independent.* This results in convolution formulas for the end–to–end delay distribution approximation. The result on the end–to–end delay in corollary 6.4.8 is therefore: In case of linear series of state independent Bernoulli servers *the assumption holds* and the convolution formula is *exact.* Clearly, compared with the decomposition approximation of Bruneel and Kim, formula (6.20) suffers from the fact that the class of networks dealt with is much more narrow.

The approach of Bruneel and Kim can be viewed as typical for dealing with more general cases. Corollary 6.4.8 contributes to the discussion in that we have identified some fundamental networks where their approximation is exact.

6.4.3 Open Tandems of Unreliable Bernoulli Servers

Consider the open tandem from section 6.4.1 and assume for simplicity that customers are indistinguishable. Then the joint queue length process $X=(X(t):t\in\mathbb{N})$ is Markov with state space $\mathbb{N}^J$.

The servers are assumed to be unreliable in the following sense: if at time t node j is in *up status* $=0$ then $\alpha_j(0,1)$ is the probability that in the present time slot the node fails, i.e., turns into *down status* $=1$ at time $t+1$. The node undergoes repair and if at time t node j is in *down status* $=1$ then $\alpha_j(1,0)$ is the probability that in the present time slot the node will be repaired, i.e., turns into *up status* $=0$ at time

$t+1$. Break down and repair events are independent over the nodes and depend on the local state of the respective node only. Customers at nodes in down status stay on there waiting for the repair of the server, customers arriving at such nodes are not allowed to enter but skip over failed servers to the next working node, possibly leaving the network, if no node in up status is in front of them. For a Markovian description of the system we supplement X and the states in $\mathbb{N}^J$ by a local supplementary variable $y_j \in \{0,1\}$ which indicates the *availability status* of the node $j = 1,\ldots,J$. We denote the supplemented Markov chain by (X,Y) with state space $E = (\mathbb{N} \times \{0,1\})^J$. $(X(t),Y(t)) = (X_1(t),Y_1(t),\ldots,X_J(t),Y_J(t)) = (n_1,y_1,\ldots,n_J,y_J)$ indicates that at time t there are n_j customers present at node j and that the availability status of that node is y_j.
Denote by $\oplus$ coordinatewise addition modulo 1 in $\{0,1\}$.

Theorem 6.4.9 *[MD06b] If (X,Y) is ergodic, then with norming constant $K < \infty$ its unique steady state distribution is*

$$\begin{aligned} &\pi((n_1,y_1,\ldots,n_J,y_J)) \qquad\qquad\qquad\qquad (n_1,y_1,\ldots,n_J,y_J) \in E, \\ &= \frac{1}{K}\left(\prod_{k=1}^{n_1+\cdots+n_J} \frac{b(k-1)}{c(k)}\right)\left(\prod_{j=1}^{J}\prod_{k=1}^{n_j} \frac{q_j(k-1)}{p_j(k)}\right)\left(\prod_{j=1}^{J} \alpha_j(y_j \oplus 1, y_j)\right) . \end{aligned}$$

6.5 Networks with Doubly Stochastic and Geometrical Servers

The doubly stochastic server (section 6.5.2) was introduced by Schassberger [Sch81] as a discrete time analogue of Kelly's *symmetric server* [Kel79], which is a generalization of the nodes in the BCMP networks in continuous time [BCMP75]: Nodes with processor sharing or Last–Come–First–Served–preemptive resume discipline or infinite servers. Exponential servers under FCFS are further building blocks of BCMP networks, generalized by Kelly to *general exponential nodes*. The discrete time analogue of these nodes are geometrical nodes (section 6.5.3).

For Kelly's networks of *general exponential* and *symmetric servers* steady state probabilities can be explicitly given in simple terms as in the BCMP case and main performance quantities can be computed. An appealing property is that first order mean values of relevant performance measures are insensitive: They remain invariant under variation of the service time distribution at symmetric servers as long as the mean service time remains invariant. Similar properties will be proved for the discrete time counterparts, see section 6.5.4.

6.5.1 Common Properties of Doubly Stochastic and Geometrical Server

Prerequisits for construction of doubly stochastic and geometrical servers are as follows: The server (node) consists of an unlimited sequence of service and waiting positions 1,2,3,..., which is controlled according to the *shift–protocol.* Whenever there are n customers present, $n \geq 1$, they occupy positions $1,\dots,n$. If the customer in position $i \in \{1,\dots,n\}$ departs then the gap is closed by shifting the customers from positions $i+1,\dots,n$ one step down into positions $i,\dots,n-1$, leaving their order invariant. If an additional customer is inserted into position $i \in \{1,\dots,n\}$, the customers previously on positions $i,\dots,n$ are shifted one step up into positions $i+1,\dots,n+1$, leaving their order invariant.

If $n > 0$ customers are present at the node service is provided to customers staying on positions $1,\dots,C(n)$, where $C(n) > 0$ is a node specific service parameter. We call positions $1,\dots,C(n)$ *busy,* while positions $C(n)+1,\dots,n$ are said to be *idle.* We set $C(0) := 0$.

If at time $t \in \mathbb{N}$ there are n customers present then there will be no arrival with probability $c(n) = (1-b(n))$ – or there is exactly one arrival which is of type m with probability $b(n) \cdot a(m) > 0$, $m \in M$, such that $\sum_{m \in M} a(m) = 1$ holds, M being a countable set of types.

Occurrence of arrivals depends on the history of the system only through the actual total population size in system, type selections are independent of the system's history. We assume late arrivals and for multiple events we assume the D/A rule in force (departure before arrival), see Figure 6.1. The state of the system is recorded at times $t \in \mathbb{N}$, just after possible departures and arrivals have happened.

6.5.2 The Doubly Stochastic Server

The amount of service a customer requests for depends deterministically on the customer's type. A customer of type $m \in M$ will request for an amount of $K(m)$ time units of service time, $K(m) \in \mathbb{N}_+$. For a discussion of this assumption and the modeling principles behind, see section 6.5.4, page 297. To exclude trivialities we assume that there exists at least one customer type who requests for more than one time unit of service time.

The state space S of the doubly stochastic server contains states x as follows:
$x := e$ for the empty node, and sequences
$x := [m(n),k(n);\dots;m(1),k(1)], \quad n \geq 1$
where n is the number of customers present at the node (*queue length*), $m(i)$ is the type of the customer on position i, and $k(i)$ his residual request for service time (*residual work*).

The service discipline is now described in a three-step procedure, where throughout $x := [m(n),k(n);\dots;m(1),k(1)]$ or $x = e$ is a generic state. In **(I)** we describe arrivals and service mechanisms and handling of multiple events in general. In **(II)** we add the feature of rearranging customers in the course of a one-step transition of the network, and in **(III)** we put some constraints on the rearrangements and describe how these rearrangements constraints resemble properties of doubly stochastic Markov transition matrices.

Definition 6.5.1 (Doubly stochastic discipline)
(I) General rules:

(1) *A customer present at time t in a busy position* $i \in \{1,\ldots,C(n)\}$ *obtains exactly one unit of service time until time* $t+1$.
If $k(i) > 1$ *then at time* $(t+1)-$ *this residual workload is diminished to* $k(i)-1$.
If $k(i) = 1$ *then this customer departs from the node at time* $(t+1)-$. *(Unless the restriction on the departure rules below in* **(3)** *are in force.)*

(2) *Assume a customer of type m observes just before his entrance (which will happen between t− and t) the node in state x.*
If $x = e$, *then the state changes to* $[m,K(m)]$;
else if $x := [m(n),k(n);\ldots;m(1),k(1)]$ *is the state of the system after at time t− all residual service times of customers on busy positions are decreased, then with probability* $1/C(n+1)$ *the state changes to*
$x := [m(n),k(n);\ldots;m(i+1),k(i+1);m,K(m);m(i-1),k(i-1);\ldots;m(1),k(1)]$,
for some $i \in \{1,\ldots,C(n+1)\}$. *A new arrival is inserted randomly into a busy position and has* immediate access *to service.*

(3) *The rule to handle these multiple events composed of arrivals and/or several departures resembles* rejection blocking *or* repetitive service *in multiple access transmission systems with limited buffer capacity [Per90], p. 455, i.e., not all of the requested transitions are allowed:*

Customers wishing jointly to depart due to their service completions have to stay at their present position for obtaining another service (retrial of transmission). This request for service time is identical to the previous service there. The only exception is:

If (several) services expire jointly and at the same time instant an arrival occurs then from the departure candidates we select randomly one who departs and is substituted at his position by the new arrival.

Therefore one single external arrival and at most one departure from the node can be observed. Especially no access conflicts can happen.

(4) *Assume that according to* **(1)** *or* **(2)** *one arrival or service completion or according to* **(3)** *a multiple event appeared and is handled. Then all the stay–on customers, i.e., those customers staying before on idle positions and customers on busy positions whose service did not expire, are permuted on their positions according to some probability law. This law may depend on what has happened in* **(1)**,**(2)**,**(3)** *and on the state of the node, in a way to be described now in detail.*

(II) Detailed rules:

In the following for state $x \in S$ *we shall call customers to be* stay–on customers *if they either occupy an idle position or if they are on a busy position* $i \le C(n)$ *showing* *a residual work of $k(i) > 1$. A customer is a* departure candidate *if she is in a busy position i with residual work $k(i) = 1$. A customer on position i is called a* fresh customer *if she shows her total service request $k(i) = K(m(i))$ as residual work. Recall: n is the queue length.*

(A) *If the state at time t is $x \neq e$, and if there is no departure candidate and no arrival occurs at time $t+1-$, then the set $A(x)$ of possible successor states at time $t+1$ of x is obtained as follows:*

Decrease the residual work of customers in busy positions by one and then permute the positions of the customers according to any permutation resulting in state $y \in A(x)$. This happens with probability

$$c(n)d(x,y) \geq 0, \qquad \sum_{y \in A(x)} d(x,y) = 1.$$

(B) *If the state at time t is x, and there is no departure candidate, and an arrival of type $m \in M$ occurs at time $t+1-$, then the set $A_{m,i}(x)$ of possible successor states of x with the new arrival inserted in position i, $i \in 1,\dots,C(n+1)$, at time $t+1$ is obtained as follows:*

Decrease the residual work of customers in busy positions by one, insert the new arrival in position i and then permute the stay–on customers on positions $1,\dots,i-1,i+1,\dots,n+1$ according to any permutation resulting in state $y \in A_{m,i}(x)$. This happens with probability

$$b(n)a(m)C(n+1)^{-1}d^{+}(x,y) \geq 0, \qquad \sum_{y \in A_{m,i}(x)} d^{+}(x,y) = 1.$$

(C) *If the state at time t is $x \neq e$, and if there is exactly one departure candidate and no arrival occurs at time $t+1-$, then the set $A(x)$ of possible successor states at time $t+1$ of x is obtained as follows:*

Decrease the residual work of customers in busy positions by one, delete the departure candidate and then permute on positions $1,\dots,n-1$ the stay–on customers according to any permutation resulting in state $y \in A(x)$. This happens with probability

$$c(n)d^{-}(x,y) \geq 0, \qquad \sum_{y \in A(x)} d^{-}(x,y) = 1.$$

(D) *If the state at time t is $x \neq e$, and if there are $k \geq 1$ departure candidates on positions $i_1,\dots,i_k$, and an arrival of type $m \in M$ occurs at time $t+1-$, then the set $A_{m,i}(x)$ of possible successor states of x at time $t+1$ with the new arrival staying in i, $i \in 1,\dots,C(n)+1$, is obtained as follows:*

Decrease the residual work of customers in busy positions by one, select at random one of the departure candidates who is allowed to depart and the new arrival is inserted in his previous position i. All other departure candidates stay on their positions and become fresh jobs again requesting for a further service of $K(m(i_l))$ time units, $l \in \{1,\dots,k\} - \{i\}$. Then permute the stay–on customers on positions $\{1,\dots,n\} - \{i_1,\dots,i_k\}$ according to any permutation resulting in state $y \in A_{m,i}(x)$. This happens with probability

$$b(n)a(m)k^{-1}d^{+}(x,y) \geq 0, \qquad \sum_{y\in A_{m,i}(x)} d^{+}(x,y) = 1.$$

(E) *If the state at time t is $x \neq e$, and if there are $k \geq 2$ departure candidates occupying positions $i_1,\ldots,i_k$, and no arrival occurs at time $t+1-$, then the set $A(x)$ of possible successor states at time $t+1$ of x is obtained as follows:*

Decrease the residual work of stay–on customers in busy positions by one, convert the departure candidates into fresh jobs staying on their previous positions and requesting for a further service of $K(m(i_l))$ time units, $l = 1,\ldots,k$. Then permute the stay–on customers on positions $\{1,\ldots,\ldots,n\} - \{i_1,\ldots,i_k\}$ according to any permutation resulting in state $y \in A(x)$. This happens with probability

$$c(n)d(x,y) \geq 0, \qquad \sum_{y\in A(x)} d(x,y) = 1\ .$$

(III) The doubly stochastic property

The doubly stochastic *property refers to the transition density matrices d, d^{+}, d^{-} introduced in* **(II)** *as we shall roughly explain the principle in case of d. In the situation described in* **(A)** *d can be considered as a transition matrix from the set of all states x other than e which do not show a departure candidate, into the union of all set of successor states $A(x)$. This matrix, described in* **(A)** *is row-stochastic. In general it is not a square matrix.*
In **(i)** *below we require that this matrix is column-stochastic as well.*
In a similar way the other cases can be interpreted.

Assume that state x shows no fresh jobs, *i.e., $k(i) < K(m(i))$ for all i.*
(i) *One type of predecessor states $y \in A^0(x)$ are those states which are obtained from x by rearranging the customers on positions $1,\ldots,n$ and increasing the residual work by one for those customers staying now on the busy positions.*
For $d(y,x)$ from **(A)** *must hold $\sum_{y\in A^0(x)} d(y,x) = 1$.*
(ii) *The second type of predecessor states $y \in A^0_{ri}(x)$ of x are those states which are obtained from x by inserting a fresh customer of type $r \in M$ in position i, $i \in \{1,\ldots,C(n+1)\}$, according to the shift protocol and then arbitrarily rearranging the customers on positions $1,\ldots,i-1,i+1,\ldots,n$ and increasing the residual work by one for those customers staying now on the busy positions.*
For $d^{-}(y,x)$ from **(C)** *must hold $\sum_{y\in A^0_{ri}(x)} d^{-}(y,x) = 1$.*

Next consider a state $x = [m(n),k(n);\ldots;m(1),k(1)]$ with exactly one fresh job *which is of type r in position i, i.e., $m(i) = r$ and $k(i) = K(r)$, and for all other positions $j \neq i$ we have $k(j) < K(m(j))$.*
(iii) *One type of predecessor states $y \in A^0(x)$ are those states which are obtained from x by deleting the fresh customer and rearranging the other customers on positions $1,\ldots,n-1$ and increasing the residual work by one for those customers staying now on the busy positions.*
For $d^{+}(y,x)$ from **(B)** *must hold $\sum_{y\in A^0(x)} d^{+}(y,x) = 1$.*
(iv) *The second type of predecessor states $y \in A^0_{r'i}(x)$ of x are those states which are*

obtained from x by deleting the fresh customer of type r on position i and substituting him by a departure candidate of type $r' \in M$ *and then rearranging the other customers on positions* $1,\dots,i-1,i+1,\dots,n$*, increasing the residual work by one for those customers staying now on the busy positions.*
For $d^{+}(y,x)$ *from* **(D)** *must hold* $\sum_{y\in A^{0}_{r'i}(x)} d^{+}(y,x) = 1$.

The remaining states are of the form $x = [m(n),k(n);\dots;m(1),k(1)]$ **with exactly** $k \geq 2$ **fresh customers** *in positions* $i_1,\dots,i_k$*, being of type* $m(i_l)$ *in position* i_l *with* $k(i_l) = K(m(i_l))$*, and for all other positions* $j \neq i_1,\dots,i_k$ *we have* $k(j) < K(m(j))$*.*
(v) *One type of predecessor states* $y \in A^{0}(x)$ *are those states which are obtained from x by fixing the fresh customers on their positions with maximum residual work and rearranging the other customers on positions* $\{1,\dots,n\} - \{i_1,\dots,i_k\}$*, increasing the residual work by one for those customers staying now on the busy positions.*
For $d(y,x)$ *from* **(E)** *must hold* $\sum_{y\in A^{0}(x)} d(y,x) = 1$.
(vi) *The second type of predecessor states* $y \in A^{0}_{r'i_l}(x)$ *of x are those states which are obtained from x by deleting the fresh customer of type* $m(i_l)$ *on position* i_l *and substituting him by a departure candidate of type* $r' \in M$*, fixing the residual fresh customers on their positions with maximal residual work and then rearranging the other customers on positions* $\{1,\dots,n\} - \{i_1,\dots,i_k\}$*, increasing the residual work by one for customers staying now on the busy positions.*
For $d^{+}(y,x)$ *from* **(D)** *must hold* $\sum_{y\in A^{0}_{r'i_l}(x)} d^{+}(y,x) = 1$.

Example 6.5.2 (Doubly stochastic disciplines) *The class of* doubly stochastic nodes *comprises especially nodes with the following queueing disciplines:* Last–Come–First–Served (preemptive resume), infinite server, random service allocation, round–robin with a preemptive modification for new arrivals *[DS81]. These disciplines are prototypes for so called* permutation queues*, where service is provided on a time shared basis, for a more in depth description see [Yat94] and [Yat90]. Permutations are special cases of the doubly stochastic reorganization rules.*

First–Come–First–Served *is not included because immediate service must be guaranteed for a* doubly stochastic *node [Sch81].*

The permutations in the doubly stochastic disciplines are in general not considered as a rule for physically moving stay–on customers. The interpretation is that service capacity of the node is redistributed to the customers. E.g., applying a suitable permutation rule would guarantee fairness of service. This was discussed for the case of *processor sharing* by Kleinrock [Kle76], pp. 166–172. Processor sharing is fair without applying a permutation rule.

Permutation rules for general symmetric servers were introduced by Yashkov [Yas80] and open the possibility to control the service nearly continuously.

The queueing discipline and the stochastic assumptions put on the systems ensure that the system can be described by a Markov chain $X = (X(t) : t \in I\!N)$ with state space S. Its steady state is of product form.

Theorem 6.5.3 (Steady state) *A steady state π of X exists if and only if $A = \sum_{n=0}^{\infty} \hat{\pi}(n) < \infty$ holds, where $\hat{\pi}(0) = c(0)^{-1}$ and for $n > 0$*

$$\hat{\pi}(n) = \left(\frac{\prod_{h=0}^{n-1} b(h)}{\prod_{h=0}^{n} c(h)} \right) \cdot (\mu - 1)^{n-C(n)} \cdot \mu^{C(n)} \cdot \prod_{i=1}^{n} C(i)^{-1}.$$

$\mu = \sum_{m\in M} a(m)K(m)$ is the mean service request of an average (typical) customer. If $A < \infty$ then the steady state of the doubly stochastic node is

$$\pi([m(n),k(n);\ldots;m(1),k(1)]) = \frac{\prod_{h=0}^{n-1} b(h)}{\prod_{h=0}^{n} c(h)} \prod_{h=1}^{n} \frac{a(m(h))}{C(h)} A^{-1}, \qquad (6.21)$$
$$[m(n),k(n);\ldots;m(1),k(1)] \in S.$$

The equilibrium queue length distribution is $\pi(n) = \hat{\pi}(n) \cdot A^{-1}$ on $\mathbb{N}$ and is insensitive under perturbations of the service time distributions as long as their mean is fixed.

6.5.3 The Geometrical Server

The amount of service a customer requests for is geometrically distributed with parameter $p \in (0,1)$ for all customers. Service times are independent and independent from the arrival process.

The state space S for the geometric node consists of elements x as follows:
$x := e$ for the empty node, and sequences
$x := [m(n);\ldots;m(1)], \quad n \geq 1$
where n is the number of customers present at the node (*queue length of the node*), $m(i)$ is the type of the customer on position i.

Definition 6.5.4 (Geometrical discipline) **(1)** *A customer present at time t in a busy position $i \in \{1,\ldots,C(n)\}$ obtains exactly one unit of service time until time $t+1$. With probability p her service ends at the end of $[t,t+1)$, and she departs from the node. (Unless the restrictions on departure rules in* **(3)** *and* **(4)** *are in force.) With probability $q = 1 - p$ she requests for at least one more service quantum.*

(2) *An arriving customer observing n other customers present enters position $n+1, n \geq 0$. (Unless the restrictions on departure rules in* **(3)** *and* **(4)** *are in force.) Applying the shift–protocol after the departure from a busy position then leads to a FCFS–based service.*

(3) *If more than one customer complete their service at the same time instant they are not allowed to depart jointly. They have to stay at their present position for obtaining just another service (retrial of transmission).*

(4) *If an arrival occurs and at one or more services expire jointly at the same time instant then the departure candidates stay for another service on their positions and the arrival candidate is rejected and lost.*

Therefore either one single external arrival or at most one service completion with a subsequent departure from the node can be observed. The time evolution of the system is described by a Markov chain $X = (X(t) : t \in \mathbb{N})$ with state space S. Its steady state π is of product form.

Theorem 6.5.5 (Steady state) *A steady state π of X exists if and only if $A = \sum_{n=0}^{\infty} \hat{\pi}(n) < \infty$ holds, where $\hat{\pi}_j(0) = c(0)^{-1}$ and for $n > 0$*

$$\hat{\pi}(n) = \left(\frac{\prod_{h=0}^{n-1} b(h)}{\prod_{h=0}^{n} c(h)} \right)^{n} \prod_{i=1}^{n} C(i)^{-1} \left(\frac{q}{p} \right)^{n} \left(\frac{1}{q} \right)^{C(n)} .$$

If $A < \infty$ holds then the steady state of X is

$$\pi([m(n);\ldots;m(1)]) = \frac{\prod_{h=0}^{n-1} b(h)}{\prod_{h=0}^{n} c(h)} \prod_{h=1}^{n} \frac{a(m(h))}{C(h)} \left(\frac{q}{p} \right)^{n} \left(\frac{1}{q} \right)^{C(n)} A^{-1}$$
$$[m(n);\ldots;m(1)] \in S. \tag{6.22}$$

The equilibrium queue length distribution is $\pi(n) = \hat{\pi}(n) \cdot A^{-1}$, $n \in \mathbb{N}$.

Remark 6.8 (Slotted Aloha–type protocol). The rules to regulate simultaneous events in geometrical nodes occurred first in models of transmission stations in a slotted Aloha–type communication system, [Kle76], section 5.11, [Woo94], section 6.2. If there are $C(n)$ active sources of traffic (stations), and if the end of a service in position $i \in \{1,\ldots,C(n)\}$ indicates that a message has to be transmitted over the shared medium, then this is possible if and only if exactly one service ends. If more than one service ends, and more than one message is tried to be send at the same time instant, all those transmission trials are not successful and have to be repeated. A common regime to resolve the conflicts is that the sources retry at random to repeat sending. This is just what is going on in a geometrical server according to **(3)** in definition 6.5.4.
Due to the memoryless property of the geometrical service time distribution the blocking mechanism according to repititive service is equivalent to what is known as communication blocking, [Per90], p.455.

6.5.4 Networks of Doubly Stochastic and Geometrical Nodes

The networks in this section are constructed along the lines of the BCMP networks [BCMP75] and Kelly's networks [Kel79]. These networks are now widely accepted as a versatile class of queueing networks, which simulate the behaviour of many complex systems. We substitute the symmetric and exponential servers of Kelly's networks by doubly stochastic and geometrical servers. We concentrate on open networks.

The network consists of J nodes and is fed by a Bernoulli arrival stream of customers which are of different types $m \in M$, M a countable set.

At any time $t \in \mathbb{N}$ there is either no arrival, with probability $c = 1 - b$, or there is exactly one arrival, being of type m with probability $b \cdot a(m) > 0$, $m \in M$, such that $\sum_{m \in M} a(m) = 1$ holds. The successive arrival and type decisions are an independent sequence, independent of the previous history.

The type m of an arriving customers specifies the route of this customer through the network: $W(m) = (W(m,1), W(m,2), \ldots, W(m,S(m)))$, where node $W(m,i)$ is the ith stage of m on her itinerary, and $1 \le S(m) < \infty$ is the length of his route. We assume for simplicity of presentation that $W(m,i) \neq W(m,i+1), 1 \le i < S(m)$. (For how to remove this restriction see [DS83].)

Nodes $1, \ldots, J', 0 \le J' \le J$, are *doubly stochastic* (section 6.5.2). Nodes $J' + 1, \ldots, J$ are geometrical nodes (section 6.5.3).

If $n_j > 0$ customers are present at node j service is provided to those customers staying on positions $1, \ldots, C(j, n_j)$, where $C(j, n_j) > 0$ is a node specific service parameter. We call positions $1, \ldots, C(j, n_j)$ *busy*, while positions $C(j, n_j) + 1, \ldots, n$ are said to be *idle*. ($C(j,0) := 0, 1 \le j \le J$.)

The amount of service time a customer requests for at a geometrical node j is geometrically distributed on $\mathbb{N}_+ = \{1, 2, \ldots\}$. A customer requests with probability $p_j(1 - p_j)^{k-1}$, $p_j \in (0,1]$, for exactly k time units of service at node j, $k \in \mathbb{N}_+$. These geometrical service times are drawn independently and independent of anything else in the history of the network.

If a customer of type $m \in M$, enters stage s, $1 \le s \le S(m)$, of her route, which is a doubly stochastic node $W(m,s) \in \{1, \ldots, J\}$, she will request for an amount of $K(m,s)$ units of service time, $K(m,s) \in \mathbb{N}_+$. We assume that for every doubly stochastic node there exists at least one customer type who requests at this node for more than one unit of service time.

The requirement of having deterministic service times at doubly stochastic nodes is not a restriction but widens considerably the possibility of modeling the probabilistic behaviour of customers on their itinerary at doubly stochastic nodes. The key is the introduction of different customer types and of the type and stage dependent behaviour: The random decision for the amount of the successive service requests at doubly stochastic nodes on a customer's route is done by selecting the customer's type when entering the network.

To be more specific: Let us assume that we have a set of customers with different (physical) customer types requesting for service according to general (type– and stage–dependent) distributions at the doubly stochastic nodes of their route. We can discriminate between different sampled sequences of requests for a specific customer type by introducing ficticious customer types. Each ficticious customer type carries information about the physical type of that customer, her routing, and her exact successive amounts of requested service at her successive stages on her itinerary. This concept even allows for using stochastically dependent service requests at the successive doubly stochastic nodes.

The local state space S_j for a doubly stochastic node $j, j \in \{1, \ldots, J'\}$ consists of elements x_j as follows:

$x_j := e_j$ for the empty node, and sequences
$x_j := [m(j,n_j),s(j,n_j),k(j,n_j);\dots;m(j,1),s(j,1),k(j,1)], \quad n_j \geq 1$
where n_j is the number of customers present at node j (*queue length at node j*), $m(j,i)$ is the type of the customer on position i, $s(j,i)$ is the stage number of this customer on his actual route, and $k(j,i)$ is his residual request for service time (his *residual work*), $1 \leq i \leq n_j$.

The local state space S_j for a geometrical node $j, j \in \{J'+1,\dots,J\}$ consists of elements x_j as follows:
$x_j := e_j$ for the empty node, and sequences
$x_j := [m(j,n_j),s(j,n_j);\dots;m(j,1),s(j,1),], \quad n_j \geq 1$
where n_j is the number of customers present at node j (*queue length at node j*), $m(j,i)$ is the type of the customer on position i, and $s(j,i)$ is the stage number of this customer on his actual route, $1 \leq i \leq n_j$.

Global states of the network are composed of these local states. The state space of the network is $S := S_1 \times S_2 \times \cdots \times S_J$ or a subset thereof.

In case of a network with general topology we have to impose some further rules which regulate the network's behaviour at instances of simultaneous events.

Definition 6.5.6 (Queueing disciplines in the network) *The customers' behaviour in the network is governed by a two-step regime. First: the arrival decision is made and at any node the customers are served individually. Second: multiple events are regulated according to the Aloha-type protocol (Remark 6.8,* repetitive service/rejection blocking*) in multiple access transmission systems, none of the requested multiple transitions is allowed:*

Customers on arrival from the outside source are lost; customers wishing to depart from the network due to a service completion at node $W(\cdot,S(\cdot))$, or wishing to enter the next stage of their route due to service completion on the present stage have to stay at their present node on their present position for obtaining just another service (retrial of transmission) .

This request for retrial service time is distributed according to the node specific geometrical distribution, if the node is geometrical. Otherwise, if the node is doubly stochastic, then the additional service request is deterministically selected and identical to the previous service there. (The additional service at the node is therefore not counted as an additional stage for the customer's passage.)

The restriction for either a single arrival or a single departure from exactly one node according to the blocking protocol (repetitive service/rejection blocking) resembles the departure protocol applied in [Miy96] for batch service disciplines in discrete time networks. At every time epoch at most one node is selected to release a batch of customers being served to be distributed over the network or partially to leave the network. As Miyazawa puts it, this model is motivated not only by the fact that it is important for discrete time queueing networks, but also by its tractability for analysis.

The network's state process $X = (X_t : t \in I\!N)$ is a Markov chain on state space $S = S_1 \times \cdots \times S_J$. S is not minimal, e.g., a customer of type $m \in M$ at a doubly stochastic node $W(m,s) \leq J'$, can show a residual work of $K(m,s)$ time units if and only if she is on a busy position of $W(m,s)$. We assume henceforth that S and $S_j, 1 \leq j \leq J$ are restricted to states with feasible workloads for the customers. A detailed description of the transition laws is given in [DS83].

We define local measures $\hat{\pi}_j = (\hat{\pi}_j(x_j) : x_j \in S_j)$ for the nodes as follows: $\hat{\pi}_j(e_j) = 1$, and if j is doubly stochastic then

$$\hat{\pi}_j([m(j,n_j),s(j,n_j),k(j,n_j);\ldots;m(j,1),s(j,1),k(j,1)]) = \left(\frac{b}{c}\right)^{n_j} \prod_{i=1}^{n_j} \frac{a(m(j,i))}{C(j,i)}$$

and if j is a geometric then

$$\hat{\pi}_j([m(j,n_j),s(j,n_j);\ldots;m(j,1),s(j,1)]) = \left(\frac{bq_j}{cp_j}\right)^{n_j} \left(\frac{1}{q_j}\right)^{C(j,n_j)} \prod_{i=1}^{n_j} \frac{a(m(j,i))}{C(j,i)}$$

Let $b_j = \sum_{(m,s):W(m,s)=j} b \cdot a(m)$, $1 \leq j \leq J$, denote the total arrival rate at node j, and μ_j the mean service request of a typical customer at node j: For a geometrical node j $\mu_j = p_j^{-1}$ and for a doubly stochastic node j $\mu_j = \sum_{(m,s):W(m,s)=j} ba(m)b_j^{-1}K(m,s)$.

Theorem 6.5.7 (Steady state) *The Markov chain $X = (X_t : t \in I\!N)$ describing the network's evolution has a steady state if and only if $A_j := \sum_{n=0}^{\infty} \hat{\pi}_j(n) < \infty$ holds for all j, where $\hat{\pi}_j(0) = 1$ and for $n > 0$*

$$\hat{\pi}_j(n) = \left(\frac{b_j(\mu_j - 1)}{c}\right)^n \cdot \left(\frac{\mu_j}{\mu_j - 1}\right)^{C(j,n)} \cdot \prod_{i=1}^{n} C(j,i)^{-1} \quad .$$

The steady state distribution is

$$\pi(x) = \prod_{j=1}^{J} \hat{\pi}_j(x_j) \cdot A_j^{-1} \quad , \qquad if \quad x = (x_1,\ldots,x_J) \in S. \tag{6.23}$$

A proof can be found in [DS83] as well as computation of performance indices.

Remark 6.9 (Related models and parallel developments).
1. Closed networks The theory for closed networks of doubly stochastic and geometrical servers is developed by Krüger [Krü83], for a summarizing description see [Dad01][section 5.6].
2. Open tandems The case of open tandems of doubly stochastic and geometrical servers allows a more detailed analysis without the restriction put on the multiple events handling, see [Sch81] or [Dad01][section 5.7]. The idea behind is a discrete time analogue of Burke's theorem [Bur56], [HB76].
3. Networks of unreliable geometrical nodes A network of geometrical nodes with a state dependent arrival stream, where the arrival probabilities depend on the total population size of the network and the arriving customer's type similar to section 6.4.1, page 285, is investigated in [MD06a]. It turns out that the steady state distri-

bution of theorem 6.5.7 is changed similarly to the expression for the arrival term in (6.17) of theorem 6.4.1.

The model is further refined in that the nodes can break down and undergo successive repair similar to the mechanisms described in Section 6.4.3. The steady state is then supplemented with a term similar to the factor concerning the breakdown and repair probabilities in theorem 6.4.9.

6.6 Batch Service and Movements Networks

Over around thirty years a set of different network models has been developed as generalization of (continuous time) product form networks with the additional features that customers are served in batches and proceed partly in different batches to other nodes or leave the network. The aim was to define the arrival, service, and routing mechanism in a way that functional descriptions of these data lead to explicit functional expressions for the steady state distributions. This was often connected with partial balance structures inside the describing Markov process and some sort of insensitivity theory. We describe in this section some prototypes of classical results to let the reader get an impression of the ideas underlying these models and some recent progress in extending the area of such models.

6.6.1 The General Network Model

The description of a rather general network model with batch services and batch movements follows Henderson and Taylor [HT90], [HT91], Miyazawa [Miy94], [Miy95], and Osawa [Osa94]. More detailed structural properties of these models can be found in [CHPT97]. (This paper contains references for further applications of the models, e.g., to Petri nets.) Parallel results are presented by Boucherie and van Dijk [BD91], [Bou92], with additional features, e.g., state dependent routing.

Consider an open network of queues with nodes numbered $1,2,\dots,J$. Customers enter the system from the outside (which is termed node 0), procede according to some routing regime through the network and eventually leave the system. The customers may be of different types which they may randomly change when entering a new node. The set M of customer types is finite.

The system evolves in discrete time $\mathbb{N}$ according to a Markovian transition law, the state process is denoted by $X = (X_t : t \in \mathbb{N})$, with state space S which is $(\mathbb{N}^{|M|})^J$ or a subset thereof. X carries the following information:
$X = (X_t : t \in \mathbb{N}) = ((X_t(j,m) : j = 1,\dots,J, m \in M) : t \in \mathbb{N})$, where $X_t(j,m) = n(j,m)$ indicates that at time t there are $n(j,m)$ customers of type m at node j. X is governed by sequences of *release vectors* $D = (D_t : t \in \mathbb{N})$ and *transformed vectors* $A = (A_t : t \in \mathbb{N})$, where for $t \in \mathbb{N}$ we have

- release vector $D_t = (D_t(0), D_t(j,m), 1 \le j \le J, m \in M)$, and

• transformed vector $A_t = (A_t(0), A_t(j,m), 1 \le j \le J, m \in M)$.
The construction of the network process is as follows:

Assume at time t the network is in state $X_t = n$. Then at time $(t+1)-$ from the source (node 0) $D_t(0) = a(0)$ customers are released for being transformed into the network. From node j there are $D_t(j,m) = a(j,m)$ customers of type m released to be transformed to other nodes or to the sink (node 0). Immediately thereafter (*between time* $(t+1)-$ *and time* $t+1$) the released customers are transformed – possibly changing their types – to their destination nodes: $A_t(0) = a'(0)$ customers depart from the network, $A_t(j,m) = a'(j,m)$ customers of type m enter node j, $1 \le j \le J, m \in M$. Updating the state of the network according to this movements and changes we obtain X_{t+1}. Formally: For

$$a = (a(0), a(j,m), 1 \le j \le J, m \in M) \text{ let } a^+ = (a(j,m), 1 \le j \le J, m \in M),$$

and similarly let A_t^+ and D_t^+ be obtained from A_t and D_t by deleting the *external* components $A_t(0)$ and $D_t(0)$. Then

$$X_{t+1} = X_t - D_t^+ + A_t^+, \quad t \in \mathbb{N}, \tag{6.24}$$

$$\text{and} \quad X_t \ge D_t^+, \qquad X_{t+1} \ge A_t^+, \tag{6.25}$$

where the inequalities are to be read coordinatewise. Having now

$$X_{t+1} = n', X_t = n, D_t = a, A_t = a' \quad then \quad n' = n - a^+ + a'^+.$$

Because new customers arrive only from the outside and vanishing customers must depart to the sink (node 0) we have the balance equation

$$D_t(0) + \sum_{j=1}^{J} \sum_{m \in M} D_t(j,m) = A_t(0) + \sum_{j=1}^{J} \sum_{m \in M} A_t(j,m). \tag{6.26}$$

The sequences A and D therefore take values from a commom state space denoted by $\mathcal{A} \subseteq \mathbb{N} \times \mathbb{N}^{J \cdot |M|}$. But the images of A and D need not be identical.

The following probabilistic assumption on A and D imply that X as given by (6.24) is a Markov chain with stationary transition probabilities:

D_t depends on the history of the system up to time t only through X_t and

$$\begin{aligned} &P(D_t = a | X_t = n, X_s = n_s, D_s = a_s, A_s = a'_s, 0 \le s < t) \qquad (6.27)\\ &= P(D_t = a | X_t = n) = q(n,a), \quad n \in S, a \in \mathcal{A}, \quad with \quad (6.25), \end{aligned}$$

A_t depends on the history of the system up to $(t+1)-$ only through D_t and

$$\begin{aligned} &P(A_t = a' | D_t = aX_s = n_s, D_s = a_s, A_s = a'_s, 0 \le s < t) \qquad (6.28)\\ &= P(A_t = a' | D_t = a) = r(a,a'), \quad a, a' \in \mathcal{A}, \quad with \quad (6.26). \end{aligned}$$

The sequence of transformations of released vectors into transformed vectors governed by (6.28) are called *routing process*. Transition probabilities of X are

$$P(X_{t+1} = n'|X_t = n) = \sum_{\substack{a,a' \in \mathcal{A} \\ n-a^+ + a'^+ = n'}} q(n,a)r(a,a'), \quad n,n' \in S.$$

From (6.26) it follows that a Markov chain on $\mathcal{A}$ determined by the routing matrix $R = (r(a,a') : a,a' \in \mathcal{A})$ is not irreducible. Vectors a and a' connected by (6.26) are assumed to be reachable vice versa via R. We assume

Assumption 6.6.1 *For transition matrix $R = (r(a,a') : a,a' \in \mathcal{A})$ on $\mathcal{A}$ denote by $\mathcal{A}_k \subseteq \mathcal{A}$ the set of states in $\mathcal{A}$ which contain exactly k customers, $k \in \mathbb{N}$. The sets $\mathcal{A}_k$ are finite and we* assume *that R restricted to each $\mathcal{A}_k$ is irreducible.*

Essential for successful analysis are the following assumptions. These or similar ones can be found in almost all papers on the subject.

Assumption 6.6.2 *There exist functions $\Phi : S \longrightarrow (0,\infty)$, $\Theta : \mathcal{A} \longrightarrow [0,\infty)$, and $\Psi : S \times \mathcal{A} \longrightarrow [0,\infty)$, such that for all $n \in S, k \in \mathbb{N}$ the partial functions*

$$\Psi(n,\cdot) : \mathcal{A}_k \longrightarrow [0,\infty), \quad a \to \Psi(n,a)$$

are constants, and

$$q(n,a) = \frac{\Psi(n-a^+,a)\cdot\Theta(a)}{\Phi(n)}, \quad n \in S, \quad a \in \mathcal{A}. \tag{6.29}$$

Assumption 6.6.3 *There exist functions $f : \mathcal{A} \longrightarrow (0,\infty), g : S \longrightarrow (0,\infty)$, such that f solves the following* traffic equations *for batch movement systems*

$$\Theta(a)f(a) = \sum_{a' \in \mathcal{A}} \Theta(a')f(a')r(a',a), \quad a \in \mathcal{A}, \tag{6.30}$$

and there is a representation

$$\frac{g(n)}{g(n-a^+ + a'^+)} = \frac{f(a)}{f(a')} \quad \forall n \in S, a,a' \in \mathcal{A} \;\; with \;\; q(n,a)r(a,a') > 0. \tag{6.31}$$

Theorem 6.6.4 (Steady state) *Suppose that assumptions 6.6.1, 6.6.2, 6.6.3 hold. Then the state process X of the network has invariant measure*

$$\hat{\pi}(n) = \Phi(n)g(n), \quad n \in S. \tag{6.32}$$

If $C = \sum_{n \in S} \Phi(n)g(n) < \infty$ then $\hat{\pi}$ can be normalized to an invariant probability

$$\pi(n) = C^{-1}\Phi(n)g(n), \quad n \in S. \tag{6.33}$$

The proof is via time reversion, see [HT90], [Miy95].

To evaluate π in (6.33) we need an explicit expression for $g(n), n \in S$, which is easier accessible than the *representation* (6.31).

For the case that in state $n \in S$ a transition to $n' = n - a^+ + a'^+$ is possible such that $a^+ - a'^+ > 0$ holds, (6.31) suggests a recursion step of the form

$$g(n) = g(n - a^+ + a'^+) \cdot \frac{f(a)}{f(a')} \quad , \tag{6.34}$$

and provided this iteration can be continued we may hope to end eventually by $g(\tilde{n})$, for some base state $\tilde{n}$, which in open networks usually can be chosen $\tilde{n} = 0$, the empty network. Provided further, that this iteration for all states can be done in a *well defined* way, we would be able to compute the network's equilibrium. *Well defined* means that the result obtained by the iterative procedure (6.34) does not depend on the path from the base state $\tilde{n}$ to other states n.

As Miyazawa [Miy94] noticed, almost all examples of open batch movement networks in the literature, where a product form equilibrium is known, show the following structure: The base state is the empty state, $\tilde{n} = 0$. Let $e_0 \in \mathcal{A}$ denote the unit vector having 1 in the 0th coordinate and other coordinates 0, and $e_{i,m} \in \mathcal{A}$ the unit vector having 1 in the (i,m)th coordinate and other coordinates 0. If $q(n,e_0)r(e_0,e_{i,m}) > 0$ then (6.31) and (6.34) imply

$$g(n + e_{i,m}^+) = g(n) \cdot \frac{f(e_{i,m})}{f(e_0)}.$$

Denoting

$$\frac{f(e_{i,m})}{f(e_0)} =: \alpha_{i,m}, \quad i = 1,\dots,J, m \in M,$$

we obtain

$$g(n) = g(0) \prod_{j=1}^{J} \prod_{m \in M} \alpha_{j,m}^{n(j,m)}, \; n \in S, \quad f(a) = f(e_0) \prod_{j=1}^{J} \prod_{m \in M} \alpha_{j,m}^{a(j,m)}, \; a \in \mathcal{A}.$$

For properties of functions obeying such a representation see [Ser93], p.149, 155.

Remark 6.10. In [HT90] and [Miy94] and the references there, a stronger condition than assumption 6.6.2 is required: $\Psi : S \times \mathcal{A} \longrightarrow [0,\infty)$, is a function of S only, independent of the second coordinate $\mathcal{A}$. In [Miy95] it was remarked that (6.6.2) suffices to prove the theorem. The interpretation is:

For $a, b \in \mathcal{A}$ we write $a \rightleftharpoons b$ if and only if a, b are members of the same communicating class with respect to transition matrix R of the routing process. Then $\Psi(\cdot,\cdot)$ depends in its second coordinate only through equivalence classes of $\rightleftharpoons$. So Ψ can be written as $(n,a) \longrightarrow \Psi(n,a) := \tilde{\Psi}(n, a/\rightleftharpoons)$, where

$$\tilde{\Psi} : S \times (\mathcal{A}/\rightleftharpoons) \longrightarrow [0,\infty)$$

is a function defined in its second coordinate on the equivalence classes of $\mathcal{A}/\rightleftharpoons$.

The introduction of the additional dependency of Ψ on a second coordinate with respect to $\mathcal{A}$ is assumed to broaden the applicability of the concept.

Remark 6.11. In definition (6.29) of the service allocation function $q(\cdot,\cdot)$ the functions Ψ, Φ, θ cannot be chosen arbitrarily, because for any n the function $q(n,\cdot)$ is a density on $\mathcal{A}$. $\Phi(n)$ is therefore the norming constant of $q(n,\cdot)$. In continuous time this restriction is not necessary because $q(\cdot,\cdot)$ is a rate.

The form of $q(\cdot,\cdot)$ allows a versatile modeling of service features. E.g., the total number of customers moving in one time step can be bounded by setting $\Theta(a) = 0$ if $a(0) + \sum_{j=1}^{J} \sum_{m \in M} a(j,m) > B$ for a prescribed bound B.

6.6.2 Walrand's S–Queues and Networks

Walrand [Wal83] introduced S-queues and their networks. He proved that driven by an arrival sequence of independent Poisson–distributed customer batches the isolated S-queues are quasi–reversible. From this it follows that these nodes can be used as building blocks of networks with product form equilibrium. Because of their structural simplicity these models were used as a versatile tool in modeling information networks and investigating their performance analysis [Woo94]. For further detailed investigation of S-queues see [CMP99]. Walrand's networks of S–queues are standard examples in the literature on service systems with batch arrivals and batch services, see [Bou92], [BD90a], and [BD91].

Our presentation of the S-queue deviates from the original one in that observation instants are slightly shifted to fit it into the framework of Section 6.6.1.

Theorem 6.6.5 *[Wal83] Let $A = (A_t : t \in \mathbb{N})$ be an $\mathbb{N}$–valued i.i.d. arrival sequence of Poisson–λ variables (the transformed variables) and $D = (D_t : t \in \mathbb{N})$ an $\mathbb{N}$–valued sequence of release variables (to the outside). With a suitable initial value X_0 the queue length sequence*

$$X_{t+1} = X_t - D_t + A_t, \quad t \in \mathbb{N},$$

defines an S–queue if for $0 \le u \le v$

$$P(D_t = u | X_s, s \le t; D_s, s < t; A_s, s \ge 0, X_t + A_t = v) = S(v,u),$$

holds for all $t \in \mathbb{N}$. If for all $v \ge 0$

$$S(v,0) = c(v), \quad \textit{and for} \quad u \ge 0$$

$$S(v,u) = \frac{c(v)}{u!}\alpha(v)\alpha(v-1)\cdots\alpha(v-u+1), \quad 0 < u \le v, \tag{6.35}$$

where $\alpha(0) = 1$ and $\alpha(u) > 0$ for $u > 0$, and $c(v)$ is such that $S(v,\cdot)$ is a density on $\{0,1,\ldots,v\}$, then $X = (X_t : t \in \mathbb{N})$ has invariant measure

$$\hat{\pi}(v) = c\frac{\lambda^v}{\alpha(0)\cdots\alpha(v)}, \quad v \ge 0.$$

If $\hat{\pi}$ can be normalized to a probability law π, then in steady state the departure sequence $D = (D_t : t \in \mathbb{N})$ is an i.i.d. sequence of Poisson–λ variables, and for all t is $(D_s : s \leq t-1)$ independent of X_t. (Quasi–reversibility for the S–queue.)

To prove product form equilibria for networks of S–queues by quasi–reversibility the network is observed at time instances when customers are released from the nodes and from the outside but are not yet deposited at the destination nodes, see lemma 12.8 in [CMP99]. But, as Henderson and Taylor remarked (see [HT90], section 3.3), these networks fit into the formalism of section 6.6.1.

This is because Walrand assumed independent Poisson–γ_j arrival sequences, $j = 1,\dots,J$, ($\gamma = \sum_{j=1}^{J} \gamma_j$), Markovian routing, and because the functions $S(\cdot,\cdot)$ can be suitably reproduced. Taking Walrand's routing probabilities $(r(i,j) : 1 \leq i,j \leq J)$ and additionally $(r(0,j) = \gamma_j/\gamma : 1 \leq j \leq J)$ we define

$$\Phi(n) = \prod_{i=1}^{J} c(n_i)^{-1} \prod_{h=1}^{n_i} \alpha_i(h)^{-1}, \quad n \in \mathbb{N}^J,$$

$$\Psi(n-a^+) = \prod_{i=1}^{J} \prod_{h=1}^{n_i-a_i} \alpha_i(h)^{-1}, \quad n \in \mathbb{N}^J, a \in \mathcal{A}, n-a \geq 0,$$

$$\Theta(a) = \frac{e^{\gamma} \gamma^{a(0)}}{\prod_{i=1}^{J} a(i)!}, \quad a \in \mathcal{A},$$

and obtain for a stable system an equilibrium probability

$$\pi(n) = C\Phi(n) \prod_{i=1}^{J} \eta(i)^{n_i}, \quad n \in \mathbb{N}^J, \tag{6.36}$$

where C is the normalizing constant and $(\eta(j) : 1 \leq j \leq J)$ is the solution of the standard traffic equation.

The steady state probabilities (6.36) are obtained in ([HT90]) and differ from that obtained by Walrand, due to different observation time points.

Closely related are the networks investigated by Woodward in [Woo96], [Woo00]. In this work emphasis is put on applying the results to modeling of ATM networks.

6.6.3 Closed Networks of Unreliable S–Queues

Generalized S–queues combine the departure rules (6.29) from assumption 6.6.2 and (6.35) of theorem 6.6.5. There are J nodes and K indistinguishable customers, the joint queue length process X is Markov on $S(K,J) = \{(x_1,\dots,x_J) \in \mathbb{N}^J : x_1 + \cdots + x_J = K\}$ with local departure probabilities

$$q_i(n_i,a_i) = \frac{\Psi_i(n_i - a_i)}{(a_i!)\Phi_i(n_i)}.$$

where $\Phi_i : \mathbb{N} \longrightarrow (0,\infty), \Psi_i : \mathbb{N} \longrightarrow [0,\infty)$, are general functions similar to those in assumption 6.6.2. Assuming independence over nodes the global departure probabilities are

$$q(n,a) := \prod_{i=1}^{J} q(n_i,a_i) = \prod_{i=1}^{N} \frac{\Psi_i(n_i - a_i)}{(a_i!)\Phi_i(n_i)}. \tag{6.37}$$

Movements are independent over customers according to irreducible routing probabilities $(r(i,j) : 1 \le i,j \le J)$ with probability solution $(\eta(j) : 1 \le j \le J)$ of the standard traffic equation.

The nodes are unreliable, i.e. can break down in the course of a time slot and are repaired after some random time. Thus, the system state has to record the availability status of the nodes. Let $\bar{I} \subseteq \{1,\ldots,J\}$ be the set of nodes under repair. The states of the network are of the form

$$(n,\bar{I}) = ((n_1,\ldots,n_N),\bar{I}) : n = (n_1,\ldots,n_J) \in S(K,J), \bar{I} \subseteq \{1,\ldots,N\}),$$

Changes in the state of the system occur due to
a) breakdowns of active nodes and/or repairs of inactive nodes, and
b) departures of customers from nodes and their arrival to other nodes.
Breakdowns and repairs are assumed to occur independently from the queue-lengths at the various nodes of the network. If, at the beginning of a time slot, the nodes in $\bar{I} \subseteq \{1,\ldots,N\}$ are inactive, the probability that, by the end of the same time slot, the set of inactive nodes is $\bar{H} \subseteq \{1,\ldots,N\}$, is given by $\gamma(\bar{I},\bar{H})$, γ being a reversible transition matrix on $\mathcal{P}\{1,\ldots,J\}$.

As long as $\bar{I} = \emptyset$, nodes are functioning and customers move as determined by these regimes. If, however, $\bar{I} \neq \emptyset$, that is, if at least one node is inactive, either all nodes immediately stop serving customers ("stalling"), or exactly the nodes in $\bar{I}$ interrupt services and reject new arrivals. Consequently, some re-routing strategy has to be applied, because the active nodes continue their services. In either case, though, all customers in service at a node that breaks down have to stay there until the node is repaired and their service time is terminated.

If the network's service is not stalled during some nodes being under repair, a customer trying to make a prohibited movement
• either is sent back to the node she has just left and is served there once more (Repeated Service – Random Destination: RS-RD),
• or she makes a virtual jump to the node of her choice; having arrived there, she immediately jumps to the next node according to r as though she had just left the respective inactive node, and so on, until she reaches an active node (skipping).

Instead of re-routing individual customers, one can also determine whether an arrival vector is permitted or prohibited and then, if the vector is prohibited, retransform it according to some global RS-RD or skipping rules, for more details see [Tip06].

Theorem 6.6.6 *[Tip06] In closed networks of generalized S-queues with departure probabilities in the form* (6.37) *and with unreliable servers the joint availability and queue length process is Markov. In case of RS-RD we assume that the routing of* *customers is reversible. The equilibrium distribution is independent of the rerouting strategies described above*

$$\pi((n_1,\dots,n_N),\bar{I}) = \tilde{G}_{K,J}^{-1}\bar{\pi}(\bar{I})\prod_{i=1}^{J}\eta(i)^{n_i}\Phi_i(n_i),$$

with normalization $\tilde{G}_{K,J}^{-1}$ *and* $\eta(\cdot)$ *the probability solution of the standard traffic equation.* $\bar{\pi}(\cdot)$ *is the probability solution of the availability balance equations*

$$\bar{\pi}(\bar{I}) = \sum_{i=1}^{N}\bar{\pi}(\bar{J})\gamma(\bar{J},\bar{I})$$

6.6.4 Networks with Triggered Batch Movements

The networks' description follows Henderson, Northcote, and Taylor [HNT95] who constructed a unification of the batch movements networks of section 6.6.1 and the recently introduced networks where events (e.g. service completions or arrivals) may trigger other events to happen. This happens in a way, that by a triggering event (e.g.) a sudden departure of specific customers may be enforced without those customers having obtained the full requested service time. A related concept is that of *negative customers*, see [Gel91] and [CP93]. For a discussion of further related models and for more references see in [CMP99] the reference note 9.8, for generalizations of the network of section 6.6.1 see chapters 4 through 11 there. Further extension are by Serfozo and Yang [SY98], and Peterson [Pet00]. All these models provide us with discrete time queueing networks by simple transformations.

Consider an open network of queues with nodes $1,2,\dots,J$. Customers are undistinguishable and enter the system from the outside (node 0), procede according to some routing regime through the network and eventually leave the system. (Different customer types which may randomly change can be incorporated, see [HNT95] section 4.)

The joint queue length process $X = ((X_t(j) : j = 1,\dots,J) : t \in \mathbb{N})$ is Markov with state space $\mathbb{N}^J$, where $X_t(j) = n_j$ indicates that at time t there are n_j customers present at node j. The one–step transitions are as follows:

If the network's state at time t is $X_t = n = (n_1,\dots,n_J)$, then a batch $a = (a_1,\dots,a_J)$ of customers is served at nodes $1,\dots,J$ resp., with probability

$$q(n,a) = \frac{\Psi(n-a)\Theta(a)}{\Phi(n)}.$$

With probability $p(a,a',a'')$ the released batch a attempts to trigger a batch $a' = (a'_1,\dots,a'_J)$ (enforcing these customers to immediately finish their service) and to deposit a further batch $a'' = (a''_1,\dots,a''_J)$ to the respective nodes, $\sum_{a'}\sum_{a''}p(a,a',a'') = 1$. $a,a',a'' \in \mathcal{A}$ must fulfill some feasibility conditions to guarantee consistent transitions. External arrivals are included into the deposited batch a'', the internal transitions are included in a, a' and a'', external departures are due to either a or a'. Batch a' accepts the triggering with probability $\Psi(n-a-a')/\Psi(n-a)$ and rejects it with probability $1-\Psi(n-a-a')/\Psi(n-a)$.

The state X_{t+1} of the network at time $t+1$ then is $n-a-a'+a''$ if the triggering was accepted and $n-a$ if it was rejected.

The *traffic equations* similarly defined to (6.30) take the form

$$\begin{aligned}&\Theta(a)f(a) \qquad\qquad a\in\mathcal{A}-\{0\} \qquad (6.38)\\ &=\sum_{a'\in\mathcal{A}}\sum_{a''\in\mathcal{A}}\Theta(a')f(a')[f(a'')p(a',a'',a)-f(a)p(a',a,a'')],\end{aligned}$$

The existence of a strict positive solution of (6.38) is nontrivial [HNT95][pp132-133], we take it as an *assumption* here.

Theorem 6.6.7 *[HNT95] If the traffic equation (6.38) has product solution* $f(a)=\prod_{j=1}^{J}y_j^{a_j}>0$, *then* X *has invariant measure* $\hat{\pi}(n)=\Phi(n)f(n)$, $n\in\mathbb{N}^J$.

Example 6.6.8 *An example of a telecommunication network where customers may trigger by their arrival additional resources is worked out in detail in [HNT95], section 4. This network model is then solved using the result of theorem 6.6.7 .*

References

[BB94] F. Baccelli and P. Bremaud. *Elements of Queueing Theory*. Springer, New York, 1994.

[BCMP75] F. Baskett, M. Chandy, R. Muntz, and F.G. Palacios. Open, closed and mixed networks of queues with different classes of customers. *Journal of the Association for Computing Machinery*, 22:248–260, 1975.

[BGdT98] G. Bolch, S. Greiner, H. de Meer, and K. S. Trivedi. *Queueing networks and Markov chains*. John Wiley, New York, 1998.

[Bou92] R.J. Boucherie. *Product form in queueing networks*. PhD thesis, Vrije Universiteit Amsterdam, 1992.

[BD90a] R.J. Boucherie and N.M.van Dijk. Spatial birth–death processes with multiple changes and applications to batch service networks and clustering processes. *Advances in Applied Probability*, 22:433–455, 1990.

[BD91] R.J. Boucherie and N.M.van Dijk. Product forms for queueing networks with state-dependent multiple job transitions. *Advances in Applied Probability*, 23:152–187, 1991.

[Box88] O. J. Boxma. Sojourn times in cyclic queues - the influence of the slowest server. In O.J. Boxma, P.J. Courtois, and G. Iazeolla, editors, *Computer Performance and Reliability*, pages 13–24. North - Holland, Amsterdam, 1988.

[BD90b] O. J. Boxma and H. Daduna. Sojourn times in queueing networks. In H. Takagi, editor, *Stochastic Analysis of Computer and Communication Systems*, pages 401–450, Amsterdam, 1990. IFIP, North–Holland.

[BK93] H. Bruneel and Byung G. Kim. *Discrete-Time Models for Communication Systems including ATM*. Kluwer Academic Publications, Boston, 1993.

[BSDP92] H. Bruneel, B. Steyaert, E. Desmet, and G.H. Petit. An analytical technique for the derivation of the delay performance of ATM switches with multiverser output queues. *Intern. Journ. of Digital and Analog Communication Systems*, 5:193–201, 1992.

[Bur56] P.J. Burke. The output of a queueing system. *Operations Research*, 4:699–704, 1956.

[Buz73] J.P. Buzen. Computational algorithms for closed queueing networks with exponential servers. *Communications of the ACM*, 16:527–531, 1973.

[CMP99] X. Chao, M. Miyazawa, and M. Pinedo. *Queueing Networks – Customers, Signals, and Product Form Solutions*. Wiley, Chichester, 1999.

[CP93] X. Chao and M. Pinedo. On generalized networks of queues with positive and negative arrivals. *Prob.Eng.Inf.Sci*, 7:301–334, 1993.

[CG96] M.L. Chaudhry and U.C. Gupta. Transient behaviour of the discrete time Geom/Geom/m/m Erlang loss model. In A.C. Borthakur and M.L. Choudhry, editors, *Probability Models and Statistics, A.J. Medhi Festschrift*, pages 133 – 145, New Delhi, 1996. New Age International Limited, Publishers.

[CHPT97] Coleman, J.L., Henderson, W., Pearce, C.E.M., and Taylor, P.G. A correspondence between product–form batch–movement queueing networks and single–movement networks. *Journal of Applied Probability*, 34:160–175, 1997.

[Coo90] R. B. Cooper. Queueing theory. In D. P. Heyman and M. J. Sobel, editors, *Stochastic Models*, volume 2 of *Handbooks in Operations Research and Management Science*, chapter 10, pages 469– 518. North-Holland, Amsterdam, 1990.

[Dad96] H. Daduna. The cycle time distribution in a cycle of Bernoulli servers in discrete time. *Mathematical Methods of Operations Research*, 44:295 – 332, 1996.

[Dad97a] H. Daduna. Discrete time analysis of a state dependent tandem with different customer types. In Christian Freksa, Matthias Jantzen, and Rüdiger Valk, editors, *Foundations of Computer Science, Potential - Theory - Cognition*, volume 1337 of *Lecture Notes in Computer Science*, pages 287–296. Springer, Berlin, 1997.

[Dad97b] H. Daduna. The joint distribution of sojourn times for a customer traversing an overtake-free series of queues: The discrete time case. *Queueing Systems and Their Applications*, 27:297–323, 1997.

[Dad97c] H. Daduna. Some results for steady–state and sojourn time distributions in open and closed linear networks of Bernoulli servers with state–dependent service and arrival rates. *Performance Evaluation*, 30:3–18, 1997.

[Dad01] H. Daduna. *Queueing Networks with Discrete Time Scale: Explicit Expressions for the Steady State Behavior of Discrete Time Stochastic Networks*, volume 2046 of *Lecture Notes in Computer Science*. Springer, Berlin, 2001.

[DPR03] H. Daduna, V. Pestien, and S. Ramakrishnan. Asymptotic throughput in discrete-time cyclic networks with queue-length-dependent service rate. *Comm. Statist.– Stochastik Models*, 19:483–506, 2003.

[DS81] H. Daduna and R. Schassberger. A discrete–time round–robin queue with bernoulli input and general arithmetic service time distributions. *Acta Informatica*, 15:251 –263, 1981.

[DS83] H. Daduna and R. Schassberger. Networks of queues in discrete time. *Zeitschrift fuer Operations Research ZOR*, 27:159 – 175, 1983.

[DT92] J. N. Daigle and St. C. Tang. The queue length distribution for multiserver discrete time queues with batch Markovian arrivals. *Comm.Statist.–Stochastic Models*, 8:665–683, 1992.

[Dij93] N. M. van Dijk. *Queueing Networks and Product Forms – A Systems Approach*. Wiley, Chichester, 1993.

[ETS92] M. El-Taha and S. Jr. Stidham. A filtered ASTA property. *Queueing Systems and Their Applications*, 11:211–222, 1992.

[ETS99] M. El-Taha and S. Jr. Stidham. *Sample-Path Analysis of Queueing Systems*. Kluwer Academic Publisher, Boston, 1999.

[Gel91] E. Gelenbe. Produkt form queueing networks with negative and positve customers. *Journal of Applied Probability*, 28:656–663, 1991.

[GN67] W.J. Gordon and G.F. Newell. Closed queueing networks with exponential servers. *Operations Research*, 15:254–265, 1967.

[GH92] A. Gravey and G. Hebuterne. Simultaneity in discrete–time single server queues with Bernoulli inputs. *Performance Evaluation*, 14:123–131, 1992.

[Hal83] S. Halfin. Batch delays versus customer delays. *The Bell System Technical Journal*, 62:2011–2015, 1983.

[HNS99] A. Harel, S. Namn, and J. Sturm. Simple bounds for closed queueing networks. *Queueing Systems and Their Applications*, 31:125–135, 1999.

[HNT95] W. Henderson, B.S. Northcote, and P.G. Taylor. Triggered batch movement in queueing networks. *Queueing Systems and Their Applications*, 21:125 – 141, 1995.

[HT90] W. Henderson and P.G. Taylor. Product form in queueing networks with batch arrivals and batch services. *Queueing Systems and Their Applications*, 6:71–88, 1990.

[HT91] W. Henderson and P.G. Taylor. Some new results on queueing networks with batch movements. *Journal of Applied Probability*, 28:409–421, 1991.

[HB76] J. Hsu and P.J. Burke. Behaviour of tandem buffers with geometric input and markovian output. *IEEE Transactions on Communications*, 24:358 – 361, 1976.

[Hun83a] J. J. Hunter. *Mathematical Techniques of Applied Probability*, volume I: *Discrete Time Models: Basic Theory*. Academic Press, New York, 1983.

[Hun83b] J. J. Hunter. *Mathematical Techniques of Applied Probability*, volume II: *Discrete Time Models: Techniques and Applications*. Academic Press, New York, 1983.

[IT99] F. Ishizaki and T. Takine. Loss probability in a finite discrete–time queue in terms of the steady state distribution of an infinite queue. *Queueing Systems and Their Applications*, 31:317 –326, 1999.

[Jac57] J.R. Jackson. Networks of waiting lines. *Operations Research*, 5:518–521, 1957.

[Kel76] F. Kelly. Networks of queues. *Advances in Applied Probability*, 8:416–432, 1976.

[Kel79] F. P. Kelly. *Reversibility and Stochastic Networks*. John Wiley and Sons, Chichester – New York – Brisbane – Toronto, 1979.

[KP83] F. P. Kelly and Phillip Pollett. Sojourn times in closed queueing-networks. *Advances in Applied Probability*, 15:638–653, 1983.

[KSK76] J. G. Kemeny, J. L. Snell, and A. W. Knapp. *Denumerable Markov Chaines*. Springer–Verlag, New York – Heidelberg – Berlin, 1976. Reprint of the book published in 1966 by Van Nostrand, Princeton.

[Kle64] L. Kleinrock. Analysis of a time–shared processor. *Naval Research Logistics Quarterly*, 10(II):59–73, 1964.

[Kle76] L. Kleinrock. *Queueing Theory*, volume II. John Wiley and Sons, New York, 1976.

[Krü83] M. Krüger. Geschlossene Warteschlangen–Netzwerke unter doppelt–stochastischen Bediendisziplinen. Diplomarbeit Technische Universität Berlin, Fachbereich Mathematik, 1983.

[Kum93] P. R. Kumar. Re–entrant lines. *Queueing Systems and Their Applications*, 13:87–110, 1993.

[LB96] K. Laevens and H. Bruneel. Discrete–time queueing models with feedback for input–buffered ATM switches. *Performance Evaluation*, 27,28:71–87, 1996.

[LR80] S. S. Lavenberg and M. Reiser. Stationary state probabilities at arrival instants for closed queueing networks with multiple types of customers. *Journal of Applied Probability*, 17:1048–1061, 1980.

[MMW89] A. Makowski, B. Melamed, and W. Whitt. On averages seen by arrivals in discrete time. In *IEEE Conference on Decision and Control, Vol. 28*, pages 1084–1086, Tampa, FL., 1989.

[MD04] C. Malchin and H. Daduna. On the structure of roundtrip time distributions in discrete time networks. Preprint 2004-03, Schwerpunkt Mathematische Statistik und Stochastische Prozesse, Fachbereich Mathematik der Universität Hamburg, 2004.

[MD05] C. Malchin and H. Daduna. An invariance property of sojourn times in cyclic networks. *Operations Research Letters*, 33:1–8, 2005.

[MD06a] C. Malchin and H. Daduna. Availability and performance analysis in a discrete time tandem network with product form steady state. In A. German, R.; Heindl, editor, *Proceedings of the GI/ITG Conference on Measuring, Modelling and Evaluation of Computer and Communications Systems*, pages 381 –398, Berlin, 2006. VDE-Verlag.

[MD06b] C. Malchin and H. Daduna. Discrete time queueing network with product form steady state: Availability and performance analysis in an integrated model. Preprint 2006-02, Schwerpunkt Mathematische Statistik und Stochastische Prozesse, Fachbereich Mathematik der Universität Hamburg, 2006. 40 p., submitted.

[Miy94] M. Miyazawa. On the characterisation of departure rules for discrete–time queueing networks with batch movements and its applications. *Queueing Systems and Their Applications*, 18:149–166, 1994.

[Miy95] M. Miyazawa. A note on my paper: On the characterisation of departure rules for discrete–time queueing networks with batch movements and its applications. *Queueing Systems and Their Applications*, 19:445–448, 1995.

[Miy96] M. Miyazawa. Stability of discrete–time Jackson networks with batch movements. In Paul Glasserman, Karl Sigman, and David D. Yao, editors, *Stochastic Networks: Stability and Rare Events*, volume 117 of *Lecture Notes in Statistics*, chapter 5, pages 76–93. Springer, New York, 1996.

[MT94] M. Miyazawa and H. Takagi. Editorial introduction to: Advances in discrete time queues, (Special issue of Queueing Systems ,Theory and Applications). *Queueing Systems and Their Applications*, 18:1–3, 1994.

[MT92] M. Miyazawa and Y. Takahashi. Rate conservation principle for discrete–time queues. *Queueing Systems and Their Applications*, 12:215–230, 1992.

[Osa94] H. Osawa. Quasi–reversibility of a discrete–time queue and related models. *Queueing Systems and Their Applications*, 18:133–148, 1994.

[Per90] H. G. Perros. Approximation algorithms for open queueing networks with blocking. In H. Takagi, editor, *Stochastic Analysis of Computer and Communication Systems*, pages 451–498. North-Holland, Amsterdam, 1990.

[PR94a] V. Pestien and S. Ramakrishnan. Asymptotic behavior of large discrete–time cyclic queueing networks. *Annals of Applied Probability*, 4:591 – 606, 1994.

[PR94b] V. Pestien and S. Ramakrishnan. Features of some discrete–time cyclic queueing networks. *Queueing Systems and Their Applications*, 18:117 – 132, 1994.

[PR99] V. Pestien and S. Ramakrishnan. Queue length and occupancy in discrete-time cyclic networks with several types of nodes. *Queueing Systems and Their Applications*, 31:327 – 357, 1999.

[PR02] V. Pestien and S. Ramakrishnan. Monotonicity and aymptotic queue length distribution in discrete-time networks. *Queueing Systems and Their Applications*, 40:313 – 331, 2002.

[Pet00] S. Peterson. General batch service disciplines - A product–form batch processing network with customer coalescence. *Mathematical Methods of Operations Research*, 52:79–97, 2000.

[Rei79] M. Reiser. A queueing network analysis of computer communication networks with window flow control. *IEEE Transactions in Communications*, COM-27:1199–1209, 1979.

[Rei82] M. Reiser. Performance evaluation of data communication systems. *Proceedings of the IEEE*, 70:171–196, 1982.

[RMW94] J.-F. Ren, J. W. Mark, and J.W. Wong. Performance analysis of a leaky–bucket controlled ATM multiplexer. *Performance Evaluation*, 19:73–101, 1994.

[STH98] Y. Sakai, Y. Takahashi, and T. Hasegawa. Discrete time multi–class feedback queue with vacations and close time under random order of service discipline. *Journal of the Operartions Research Society of Japan*, 41:589–609, 1998.

[Sch81] R. Schassberger. The doubly stochastic server: A time–sharing model. *Zeitschrift fuer Operations Research ZOR*, 25:179–189, 1981.

[Ser93] R. F. Serfozo. Queueing networks with dependent nodes and concurrent movements. *Queueing Systems and Their Applications*, 13:143–182, 1993.

[Ser99] R. F. Serfozo. *Introduction to Stochastic Networks*, volume 44 of *Applications of Mathematics*. Springer, New York, 1999.

[SY98] R. F. Serfozo and B. Yang. Markov network processes with string transitions. *Annals of Applied Probability*, 8:793–821, 1998.

[SM81] K. C. Sevcik and I. Mitrani. The distribution of queueing network states at input and output instants. *Journal of the Association for Computing Machinery*, 28:358–371, 1981.

[SG97] V. Sharma and N. D. Gangadhar. Some algorithms for discrete time queues with finite capacity. *Queueing Systems and Their Applications*, 25:281–305, 1997.

[SZ94] K. Sohraby and J. Zhang. Spectral decomposition approach for transient analysis of multi–server discrete–time queues. *Performance Evaluation*, 21:131–150, 1994.

[Tak93] H. Takagi. *Queueing Analysis: A Foundation of Performance Analysis*, volume 3. North–Holland, New York, 1993. Discrete-Time Systems.

[Tip06] K. Tippner. Closed networks of generalized S-queues with unreliable servers. In H.-D. Haasis, H. Kopfer, and J. Schneeberger, editors, *Operations Research Proceedings 2005*, pages 417–422, Heidelberg, 2006. Springer.

[TGBT94] P. Tran-Gia, C. Blondia, and D. Towsley. Editorial introduction to: Discrete–time models and analysis methods, (Special issue of Performance Evaluation). *Performance Evaluation*, 21:1–2, 1994.

[Wal83] J. Walrand. A discrete–time queueing network. *Journal of Applied Probability*, 20:903 – 909, 1983.

[Wal88] J. Walrand. *An introduction to queueing networks*. Prentice –Hall International Editions, Englewood Cliffs, 1988.

[Wal90] J. Walrand. Queueing networks. In D. P. Heyman and M. J. Sobel, editors, *Stochastic Models*, volume 2 of *Handbooks in Operations Research and Management Science*, chapter 11, pages 519–603. North–Holland, Amsterdam, 1990.

[Whi86] P. Whittle. *Systems in Stochastic Equilibrium*. Wiley, Chichester, 1986.

[Wol82] R.W. Wolff. Poisson arrivals see time averages. *Operations Research*, 30:223–231, 1982.

[Woo94] M.E. Woodward. *Communication and Computer Networks: Modelling with Discrete–Time Queues*. IEEE Computer Society Press, Los Alamitos, CA, 1994.

[Woo96] M.E. Woodward. Product-form closed discrete-time queueing networks with finite capacity shared buffer nodes. *Electronics Letters*, 32(20):1875–1876, 1996.

[Woo00] M.E. Woodward. Product form solutions for discrete-time queueing networks with bursty traffic. *Electronics Letters*, 36(17):1512–1514, 2000.

[Yas80] S.F. Yashkov. Properties of invariance of probabilistic models of adaptive scheduling in shared–use systems. *Automatic control and computer science*, 14:46–51, 1980.

[Yat90] R. D. Yates. *High speed round-robin queueing networks*. PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 1990.

[Yat94] R. D. Yates. Analysis of discrete time queues via the reversed process. *Queueing Systems and Their Applications*, 18:107–116, 1994.

Chapter 7
Decomposition and Aggregation in Queueing Networks

Tijs Huisman and Richard J. Boucherie

Abstract This chapter considers the decomposition and aggregation of multiclass queueing networks with state-dependent routing. Combining state-dependent generalisations of quasi-reversibility and biased local balance, sufficient conditions are obtained under which the stationary distribution of the network is of product-form. This product-form factorises into one part that describes the nodes of the network in isolation, and one part that describes the routing and the global network state. It is shown that a decomposition holds for general nodes if the input-output behaviour of these nodes is suitably compensated by the state-dependent routing function. When only a subset of the nodes is of interest, it is shown that the other nodes may be aggregated into nodes that only capture their global behaviour. The results both unify and extend existing classes of product-form networks, as is illustrated by several cases and an example of an assembly network.

7.1 Introduction

In the analysis of queueing networks, two at first sight different techniques have been used to derive product form results: quasi-reversibility and local balance. Quasi-reversibility is a property of the nodes of the network, roughly stating that they should preserve input and output flows when they are considered in isolation and fed by a Poisson process. If such nodes are coupled into a network by Markov

Tijs Huisman
ProRail, Utrecht, The Netherlands
e-mail: Tijs.Huisman@prorail.nl

Richard J. Boucherie
University of Twente, department of Applied Mathematics, Enschede, The Netherlands
e-mail: r.j.boucherie@utwente.nl

R.J. Boucherie and N.M. van Dijk (eds.), *Queueing Networks: A Fundamental Approach*, International Series in Operations Research & Management Science 154,
DOI 10.1007/978-1-4419-6472-4_7, © Springer Science+Business Media, LLC 2011

routing, the stationary distribution factorises over the nodes, i.e., is of product form (see [18, 25]). When using local balance, however, the nodes are not analysed in isolation first. Instead, the local balance equations for the entire network are considered and shown to hold in more detailed form (usually per node) under the assumed product form stationary distribution (see [5, 14]). This technique has the advantage that state-dependent routing can be analysed too.

Recently, both techniques have been combined. Boucherie [3] considers a network of quasi-reversible nodes linked by state-dependent routing. If the process associated with the routing (called the global process) satisfies local balance, the stationary distribution of the network is shown to factorise into the stationary distributions of the nodes in isolation, and the stationary distribution of the global process. Chao and Miyazawa [11] extend the definition of quasi-reversibility, allowing input and output rates of customers to differ from each other. When nodes satisfying this extended form of quasi-reversibility are coupled into a network by Markov routing, the network is shown to have a product form stationary distribution. In [12] it is demonstrated that this product form result can be proved using *biased local balance*. This is an extension of local balance allowing unbalance in the local balance equations to be compensated by a constant bias term. When the nodes are quasi-reversible with equal input and output rates, the bias terms are zero, and biased local balance reduces to ordinary local balance.

This chapter combines and extends the results of [3] and [12] to networks with more general nodes, and more general state-dependent routing. As in [3], we introduce local processes describing the nodes in isolation, and a global process describing the routing process. For the global process the definition of biased local balance of [12] is extended, allowing state-dependent bias terms. For the local processes, quasi-reversibility is further generalised to include state-dependent input rates, and a state-dependent difference between input and output rates. This difference can be interpreted as the bias of the local process with respect to the outside of a node, similar to the bias of the global process. If the bias of the nodes with respect to their outside is suitably compensated by the bias of the global process, the network allows a decomposition into the global process and the local processes. Thus, this chapter combines state-dependent generalisations of the quasi-reversibility results in [3] and of the biased local balance results in [12].

Decomposition

The first part of this chapter is concerned with the *decomposition* of queueing networks. A queueing network can be decomposed if its stationary distribution factorises into the stationary distributions of the nodes of which the network is comprised; the network is then of product form. Apart from the theoretical interest, decomposition results are also of substantial practical importance: finding the stationary distribution of an entire queueing network usually requires an enormous computational effort, whereas the stationary distribution of a single node can be found relatively easily.

The first, and perhaps most famous, decomposition results for queueing networks have been reported by Jackson [17], who considered a single class queueing network of queues with exponential service times, where customers move between the queues according to fixed routing probabilities, and arrive at the network according to a Poisson process with rate equal to the throughputs that can be obtained from the routing probabilities via the so-called traffic equations. Extensions of this result include closed queueing networks, specific service disciplines for non-exponential service times, and multiclass queueing networks, where classes differ in routing and - again under certain service disciplines - in service times, see, for example, the BCMP networks [2].

It was shown that these results were a consequence of local balance [26, 27], and later that these results were also a consequence of a special input/output property of the queues in the network, called quasi-reversibility (see, for example [18]): when a queue is considered in isolation with Poisson arrivals, the time-reversed Process describing this queue also has Poisson arrivals with the same rates as the original (time-forward) process. The two worlds of local balance and quasi-reversibility have since then moved on parallel tracks. Some product-form results, such as those for networks with blocking [5] were developed by local balance conditions, and are believed not to be available via quasi-reversibility. Other results, such as for networks with negative customers [15] were rapidly shown to be due to an extension of local balance [7]. Later, also the concept of quasi-reversibility was extended by allowing that customer classes depart from the nodes at a different rate from which they entered, which allows customers to change class in the queue, and includes negative customers, see [12]. Networks of quasi-reversible queues linked via state-dependent routing were considered in [3]. Due to the state-dependent nature of the routing, it is not possible to determine the throughput from the traffic equations. Instead, the traffic equations are replaced by a stochastic process, called the global process, that describes the number of customers in each node of the network. A decomposition of the network into the stationary distributions of the nodes and the stationary distribution of the global process is obtained under the condition that all nodes are quasi-reversible with arrival rate one, and the global process - describing the number of customers in each node, as if each node emits customers with constant rate one - satisfies local balance. Via these results, the worlds of local balance and quasi-reversibility seem to re-join the same track. This chapter provides a unified framework for quasi-reversibility and local balance.

Aggregation

The second part of this chapter is concerned with *aggregation* of queueing networks. A stochastic process is the aggregation of a queueing network with respect to an aggregation function on the state of the network, if this process describes - in probability, as well as in probability flow - the evolution of the aggregate state in the network, see [9] for a general definition.

Aggregation results are commonly referred to as Norton's theorem. Norton's theorem for queueing networks states that under certain conditions on the structure of the queueing network it is possible to replace a subset of the queueing network by a single station such that for the feature of interest (e.g. equilibrium distribution, throughput, average number of customers) the behaviour of the rest of the network remains unchanged. Norton's theorem for queueing networks was originally introduced by Chandy *et al.* [10] as an efficient aggregation method for queueing networks similar to Norton's theorem from electrical circuit theory. They prove the aggregation method to be correct for queueing networks of the BCMP-type [2] consisting of two subnetworks of which the subnetwork of interest is a single station. The results of [10] can easily be generalised to subnetworks consisting of several stations such that customers enter the subnetwork through a single input node and leave the subnetwork through a single output node. Balsamo and Iazeolla [1], Kritzinger *et al.* [19], and Vantilborgh [23] extend Norton's theorem to BCMP-networks consisting of two arbitrary subnetworks. A further extension is given by Towsley [22], where elementary state-dependent routing is incorporated. An additional extension is presented in Hsiao and Lazar [16], where it is shown that Norton's equivalent can be seen as a conditional expectation.

The relation between quasi-reversibility and Norton's theorem is introduced in Walrand [24]. Walrand considers a queueing network containing two quasi-reversible components, and shows that a quasi-reversible component may be replaced by an equivalent server. In Brandt [8] this result is extended to queueing networks of multiple quasi-reversible components linked by Markov routing, that is by state-independent routing. Pellaumail [21] shows that components of a closed network with state-dependent routing can be replaced by equivalent servers under a type of quasi-reversibility condition. Both the method and the construction of the equivalent servers require the network to be a closed network. Boucherie and van Dijk [6] discuss Norton's theorem for queueing networks consisting of product form components linked by state-dependent routing. All components can be aggregated into equivalent servers independently, and for the detailed behaviour of components it is allowed to analyse the behaviour of components as open networks in isolation (not part of the queueing network). Additional results for networks consisting of multiple components linked by state-dependent routing are reported in Van Dijk [13], where product form results for networks in which the routing probabilities depend only on the total number of customers present in the components are derived. Boucherie [3] combines the results of Boucherie and van Dijk [6] and Brandt [8]. This gives an extension of Norton's theorem to queueing networks comprised of quasi-reversible components linked by state-dependent routing. This is an extension of the results of [6] since the components in isolation are now assumed to be quasi-reversible and of [8] since the routing process is allowed to be state-dependent, such as most notably including blocking and alternative routing. A key difference with other methods is that subnetworks are analysed as open networks in isolation and not by shortcircuiting of the components. This substantially simplifies the construction of the equivalent servers.

In this chapter we extend the aggregation result of [3] to our model: we show that the global process is the aggregation of the network with respect to the global state. Moreover, we show that under some additional restrictions on the arrival rates, the local processes are also aggregations of the network with respect to the detailed state of the nodes. To obtain the necessary arrival rates for this aggregation, an iterative algorithm can be used. This algorithm appears to be similar in spirit to Marie's method [20] to compute approximations for the steady-state distribution in queueing networks with non-quasi-reversible nodes and fixed routing, and thus allows development of new approximation methods, allowing global processes that do not satisfy local balance, allowing state-dependent routing, and general global states.

Examples and outline
To make the relation with the models and assumptions of [3] and [12] more explicit, we consider them as a special case. Somewhat surprisingly, it appears that our results reduce to those of [3] if there is only one customer class, and the global state represents the number of customers in a node: the state-dependent arrival and departure rates do not lead to further extensions. This, however, only holds for single class networks. By defining a trivial global state, our model and results reduce to those of [12]. This is, in fact, almost immediate, since in this way all state-dependence is reduced. We then proceed with pull networks, in which a transition is initiated by the arrival of a customer to a queue, and subsequently a customer is removed from the originating queue [4]. Finally, we consider decomposition for assembly network.

The chapter is organised as follows. In section 7.2 the network model is described and the definitions of the global and the local processes are given. Section 7.3 presents our decomposition results, and section 7.4 our aggregation results. Examples are included in section 7.5.

7.2 Model

Consider a network comprised of N interacting nodes, labelled $n = 1, 2, \dots, N$, and an outside node, labelled node 0, in which customers of classes $\bigcup_{n=0}^{N} \{\mathcal{A}_n \cup \mathcal{D}_n\}$ route among the nodes, where $\mathcal{A}_n$ resp. $\mathcal{D}_n$ is the set of customer classes that may arrive to resp. depart from node n, $n = 0, \dots, N$. Interaction among the nodes is due to customers routing among the nodes as well as due to the state of nodes influencing the behaviour of other nodes. This interaction is specified below. First, we will describe the nodes. Then, the interaction between the nodes is characterised.

7.2.1 The nodes

Consider the state-space $\mathcal{S}_n$, with states x_n. Define the mapping $G_n : \mathcal{S}_n \to G_n(\mathcal{S}_n)$, and $X_n = G_n(x_n)$. We will refer to X_n as global state corresponding to the detailed state x_n. The global state may be seen as an aggregate state (thus containing aggregate information of the node that is of interest for its performance, such as the number of customers), but will also play a more technical role in describing the interaction between the nodes (i.e. arrival and departure processes, and the routing between the nodes). The set $G_n(\mathcal{S}_n)$ will be referred to as the global state-space of node n.

We distuinguish three types of state changes: due to an arrival, due to a departure, and due to an internal change, only. The behaviour of node n in isolation is characterised as follows, see [28] for a similar characterisation.

Definition 7.1 (Local process). Consider node n. $\mathcal{A}_n$ resp. $\mathcal{D}_n$ is the set of customer classes that may arrive resp. depart from node n. For each $c \in \mathcal{A}_n \cup \mathcal{D}_n$, let $A_n^c : G_n(\mathcal{S}_n) \rightarrow G_n(\mathcal{S}_n)$, and $D_n^c : G_n(\mathcal{S}_n) \rightarrow G_n(\mathcal{S}_n)$ $1-1$ mappings such that D_n^c is the inverse of A_n^c.

- In an arrival transition, upon arrival of a class $c \in \mathcal{A}_n$ customer at node n, the detailed state changes from $x_n \in \mathcal{S}_n$ to $x'_n \in \mathcal{S}_n$ with probability $a_n^c(x_n, x'_n)$, and the global state changes from $X_n = G_n(x_n)$ to $A_n^c(X_n)$, where $a_n^c(x_n, x'_n)$ is an honest probability function:
$$\sum_{x' \in \mathcal{S}_n} a_n^c(x_n, x'_n) = 1, \quad x_n \in \mathcal{S}_n, c \in \mathcal{A}_n. \tag{7.1}$$
- In a departure transition in detailed state x_n a state change to state x'_n causing a departure of a class $c \in \mathcal{D}_n$ customer occurs at rate $d_n^c(x_n, x'_n)$. This detailed state change results in a global state change from $X_n = G_n(x_n)$ to $D_n^c(X_n)$.
- Node n initiates internal transitions from state x_n to state x'_n with rate $i_n(x_n, x'_n)$. Internal transitions do not cause a departure or arrival and do not change the global state, i.e., $G_n(x_n) = G_n(x'_n)$.
- Consider the set of functions $\lambda_n = (\lambda_n^c : G_n(\mathcal{S}_n) \rightarrow \mathbb{R}_0^+; c \in \mathcal{A}_n)$. The local process $\mathcal{L}_n(\lambda_n)$ is the Markov chain with state-space $\mathcal{S}_n$ and transition rates $q_n(x_n, x'_n; \lambda_n)$ from state $x_n \in \mathcal{S}_n$ to state $x'_n \in \mathcal{S}_n$ defined by
$$q_n(x_n, x'_n; \lambda_n) = \sum_{c \in \mathcal{A}_n} \lambda_n^c(G_n(x_n)) a_n^c(x_n, x'_n) + \sum_{c \in \mathcal{D}_n} d_n^c(x_n, x'_n) + i_n(x_n, x'_n). \tag{7.2}$$

Observe that, upon arrival of a class c customer in state x_n, the global state changes from $X_n = G_n(x_n)$ to $X'_n = A_n^c(X_n)$, and the detailed state may change to all $x'_n \in \{x : G_n(x) = A_n^c(X_n)\}$, which also implies that $a_n^c(x_n, x'_n) = 0$ if $G_n(x'_n) \neq A_n^c(G_n(x_n))$. The detailed state may represent the detailed content of a queue, and the global state the number of customers in this queue: upon arrival of a single customer, the global state then always changes from X_n to $X_n + 1$, where the detailed state change then may reflect the position of the customer in the queue, see e.g. the (ϕ, γ, δ) protocol introduced in [18], chapter 3, to represent queue disciplines such as FIFO, LIFO and PS. A class c customer may also represent a batch of customers by defining $A_n^c(X_n) = X_n + b_n^c$, where b_n^c denotes the class c batch size arriving at node n. Moreover, b_n^c may be set to a negative value: the number of customers is then decreased upon arrival of a class c customer. Such a customer may reflect a signal in a computer network, that removes tasks at a server. In literature, such customers have also been referred to as negative customers, see e.g. [15]. Departure transitions satisfy similar conditions as arrival transitions. Upon a departure, the global state change is unique, determined solely by the current global state and the class of the departing customer, whereas the detailed state may change from x_n to all $x'_n \in \{x : G_n(x) = D_n^c(x_n)\}$, which also implies that $d_n^c(x_n,x'_n) = 0$ if $G_n(x'_n) \neq D_n^c(G_n(x_n))$. Internal transitions may correspond e.g. to completion of service phases, and - in nodes representing a subnetwork of queues - movements of customers between the queues in the subnetwork. As internal transitions do not change the global state, it must be that $i_n(x_n,x'_n) = 0$ if $G_n(x'_n) \neq G_n(x_n)$.

Remark 7.1. The class of arriving customers $\mathcal{A}_n$ is not required to coincide with the class of departing customers $\mathcal{D}_n$. As a consequence, the inverse A_n^c of D_n^c needs not be a function that corresponds to the global state change of an arriving transition, i.e., it may be that class c customers arrive to node n, but do not depart from node n. □

We assume that the local process $\mathcal{L}_n(\lambda_n)$ is ergodic. Let $\pi_n(x_n;\lambda_n)$ denote the stationary probability that $\mathcal{L}_n(\lambda_n)$ is in state x_n, i.e., for all $x_n \in \mathcal{S}_n$,

$$\sum_{x'_n \in \mathcal{S}_n} \left\{ \pi_n(x_n;\lambda_n) q(x_n,x'_n;\lambda_n) - \pi_n(x'_n;\lambda_n) q(x'_n,x_n;\lambda_n) \right\} = 0,$$

and let

$$p_n(X_n;\lambda_n) = \sum_{\{x_n : G_n(x_n) = X_n\}} \pi_n(x_n;\lambda_n), \tag{7.3}$$

denote the stationary probability that $\mathcal{L}_n(\lambda_n)$ is in global state X_n.

Observe that the transition rates (7.2) characterise the arrival rate of customers to node n via the state-dependent arrival rate functions λ_n. The arrival processes at node n can be described by a state-dependent Poisson process, whose rate $\lambda_n^c(G_n(x_n))$ is assumed to depend on the global state $X_n = G_n(x_n)$ of this node, only. For the departure process, which - in correspondence with [18, 12] - will be described by the arrival rate in the time-reversed process, a similar assumption is made.

Assumption 7.2.1 *For the local process $\mathcal{L}_n(\lambda_n)$, $c \in \mathcal{D}_n$, we assume that the arrival rate of class c customers in state x_n of the stationary time-reversed process of $\mathcal{L}_n(\lambda_n)$ depends on x_n through the global state $X_n = G_n(x_n)$, only. We will denote this rate by $\mu_n^c(X_n;\lambda_n)$:*

$$\mu_n^c(X_n;\lambda_n) = \sum_{x'_n \in \mathcal{S}_n} \frac{\pi_n(x'_n;\lambda_n)}{\pi_n(x_n;\lambda_n)} d_n^c(x'_n,x_n), \quad X_n \in G_n(S_n). \tag{7.4}$$

Quasi-reversibility plays a key-role in the theory of product form networks. Kelly [18] calls a node quasi-reversible, if, for a constant arrival rate function, the arrival rate of the time-reversed local process is constant, and equal to the arrival rate in the original (time-forward) process. This, in particular, implies that both the arrival and departure processes are Poisson processes with equal intensity, and independent of the state of a node. Chao and Miyazawa [12] have extended this definition by allowing arrival and departure rates to differ from each other: in their definition a node is quasi-reversible, if, for constant arrival rate functions, the departure process is a Poisson process that is independent of the state of a node. To distinguish these two definitions, we will call the latter form generalised quasi-reversible. We summarise the above in the following definition.

Definition 7.2 ((Generalised) quasi-reversibility). Let $\hat{\lambda}_n = (\hat{\lambda}_n^c : G_n(\mathcal{S}_n) \to \mathbb{R}_0^+; c \in \mathcal{A}_n)$ be a set of constant functions. If $\mathcal{A}_n = \mathcal{D}_n$, and, for $c \in \mathcal{D}_n$, $\mu_n^c(X_n; \hat{\lambda}_n)$ is constant in X_n and equal to $\hat{\lambda}_n^c$, then the local process $\mathcal{L}_n(\hat{\lambda}_n)$ is said to be quasi-reversible. If, for $c \in \mathcal{D}_n$, $\mu_n^c(X_n; \hat{\lambda}_n)$ is constant in X_n, then the local process is said to be generalised quasi-reversible.

In the analysis below, we do not require generalised quasi-reversibility. Instead, we use the more general form of Assumption 7.2.1, and invoke a more general form of partial balance.

7.2.2 Interaction between the nodes

Nodes are coupled via a global process. Let $X = (X_1, \dots, X_N)$ denote the global state of the network, with X_n the global state of node n. The global state-space of the network, $\mathcal{S}_g \subseteq G_1(\mathcal{S}_1) \times \dots \times G_N(\mathcal{S}_N)$, is the set of all possible global states in the network. The global state of the network affects the interaction in three ways. Routing of customers between the nodes may depend on the global state of the network, arrivals to and departure from the network may depend on the global state, and the global state of a node may cause nodes to speed up or slow down. We use the following notation. For $X \in \mathcal{S}_g$, $T_{nn'}^{cc'}(X)$ denotes the vector obtained from X, by replacing the n-th component by $D_n^c(X_n)$, and the n'-th component by $A_{n'}^{c'}(X_{n'})$, $n, n' = 0, \dots, N$, where $n = 0$, or $n' = 0$ does not result in a change of state of that component.

Definition 7.3 (Global process). Let $\mathcal{A}_0$ resp. $\mathcal{D}_0$ denote the set of customer classes that may leave resp. enter the network. Consider state $X \in \mathcal{S}_g$.

- A class $c \in \mathcal{D}_0$ customer enters the network at rate $M_0^c(X)$, and arrives at node n', $n' = 1, \dots, N$, as a class $c' \in \mathcal{A}_{n'}$ customer with probability $R_{0n'}^{cc'}(X)$. The global state changes from X to $T_{0n'}^{cc'}(X)$.
- A class $c \in \mathcal{D}_n$ customer departing from node n leaves the network as a class $c' \in \mathcal{A}_0$ customer with probability $R_{n0}^{cc'}(X)$. The global state changes from X to $T_{n0}^{cc'}(X)$.
- A class $c \in \mathcal{D}_n$ customer departing from node n, $n = 1, \dots, N$, routes to node n', $n' = 1, \dots, N$, $n' \neq n$, as a class $c' \in \mathcal{A}_{n'}$ customer with probability $R_{nn'}^{cc'}(X)$. The global state changes from X to $T_{nn'}^{cc'}(X)$.
- The rate of change of node n, $n = 1, \dots, N$, for internal and departure transitions is $N_n(X)$.
- The routing probabilities $R_{nn'}^{cc'}(X)$ are honest:

$$\sum_{n'=0,n'\neq n}^{N} \sum_{c'\in\mathcal{A}_{n'}} R_{nn'}^{cc'}(X) = 1, \;\; X \in \mathcal{S}_g, c \in \mathcal{D}_n, n = 0,\ldots,N. \tag{7.5}$$

- Consider the set of functions $M = (M_n^c : G_n(\mathcal{S}_n) \to \mathbb{R}_0^+; c \in \mathcal{D}_n, n = 1,\ldots,N)$. The global process $\mathcal{G}(M)$ is the Markov chain with state-space $\mathcal{S}_g$ and transition rates $Q(X,X';M)$ from state $X \in \mathcal{S}_g$ to state $X' \in \mathcal{S}_g$ defined by

$$Q(X,T_{nn'}^{cc'}(X);M) = \begin{cases} M_0^c(X)R_{0n'}^{cc'}(X) & n = 0, \\ M_n^c(X_n)N_n(X)R_{nn'}^{cc'}(X) & n = 1,\ldots,N, \end{cases}$$

for $n' = 0,\ldots,N$, $n' \neq n$, $c \in \mathcal{D}_n$ and $c' \in \mathcal{A}_{n'}$.

The global process describes the global state of the network, as if node n in isolation (i.e. without the multiplication factor $N_n(X)$) emits customers at rate $M_n^c(X_n)$. We will call $M_n^c(X_n)$ the nominal departure rate of class c customers from node n. The global and local processes are closely intertwined, as will become clear later. In the formulation of the global process, the nominal departure rates $M_n^c(X_n)$ depend on the local process. Furthermore, the arrival rates $\lambda_n^c(G_n(x_n))$ in the local processes depend on the global process. These relations will be made explicit when we define our network in Definition 7.4.

We assume that the global process $\mathcal{G}(M)$ is ergodic. Let $\Pi(X;M)$ denote the stationary probability that $\mathcal{G}(M)$ is in state X, i.e., for all $X \in \mathcal{G}(M)$, $c \in \mathcal{D}_n$, $n = 0,\ldots,N$,

$$\sum_{n,n'=0,\; n'\neq n}^{N} \sum_{c\in\mathcal{A}_n,\; c'\in\mathcal{A}_{n'}} \{\Pi(X;M)Q(X,T_{nn'}^{cc'}(X);M) \\ - \Pi(T_{nn'}^{cc'}(X);M)Q(T_{nn'}^{cc'}(X),X;M)\} = 0. \tag{7.6}$$

Let

$$P_n(X_n;M) = \sum_{\{X':X'_n=X_n\}} \Pi(X';M),$$

denote the marginal stationary probability that the global state of node n is X_n.

Our results are formulated via the nominal departure rates $M_n^c(X_n)$, and the departure rates of the time-reversed process that will be used to characterise the arrival processes at the nodes. Let $\Lambda_0^c(X;M)$ denote the class $c \in \mathcal{D}_0$ departure rate in the time-reversed process of $\mathcal{G}(M)$. Then

$$\Lambda_0^c(X;M) = \sum_{n'=1}^{N} \sum_{c'\in\mathcal{D}_{n'}} \frac{\Pi(T_{0n'}^{cc'}(X);M)}{\Pi(X;M)} M_{n'}^{c'}(A_{n'}^{c'}(X_{n'}))N_{n'}^{c'}(T_{0n'}^{cc'}(X))R_{n'0}^{c'c}(T_{0n'}^{cc'}(X)). \tag{7.7}$$

The nominal departure rates $M_n^c(X_n)$ of node n depend only on the global state of node n, $n = 1,\ldots,N$. We assume that this is also the case for the nominal departure rates in the time-reversed process.

Assumption 7.2.2 *For the global process $\mathcal{G}(M)$, $c \in \mathcal{A}_n$, $n = 1,\ldots,N$, and $X \in \mathcal{S}_g$, we assume that the nominal departure rate of class c customers from node n in state X of the stationary time-reversed process of $\mathcal{G}(M)$ depends on the global state X_n only. We will denote this nominal departure rate by $\Lambda_n^c(X_n;M)$:*

$$\Lambda_n^c(X_n;M)N_n(X) = \sum_{c'\in\mathcal{D}_0} \frac{\Pi(T_{n0}^{cc'}(X);M)}{\Pi(X;M)} M_0^{c'}(T_{n0}^{cc'}(X))R_{0n}^{c'c}(T_{n0}^{cc'}(X))$$

$$+\sum_{n'=1}^{N}\sum_{c'\in\mathcal{D}_{n'}} \frac{\Pi(T_{nn'}^{cc'}(X);M)}{\Pi(X;M)} M_{n'}^{c'}(A_{n'}^{c'}(X_{n'}))N_{n'}^{c'}(T_{nn'}^{cc'}(X))R_{n'n}^{c'c}(T_{nn'}^{cc'}(X)). \qquad (7.8)$$

In general, the time-reversed departure rate (7.8) will depend on the global state X of the network. The asssumption that this rate is equal to $\Lambda_n^c(X_n;M)N_n(X)$, where $\Lambda_n^c(X_n;M)$ depends on X through the global state X_n of node n, only, seems to be rather restrictive. This is not the case. Assumption 7.2.2 includes local balance, a common assumption for queueing networks with state-dependent routing. To this end, note that if $\mathcal{A}_n = \mathcal{D}_n$, and $\Lambda_n^c(X_n;M) = M_n^c(X_n)$, $X_n \in G_n(\mathcal{S}_n)$, $c \in \mathcal{A}_n$, $n = 1,\ldots,N$, then, from (7.8),

$$M_n^c(X_n)N_n(X)\Pi(X;M) = \sum_{c'\in\mathcal{D}_0} \Pi(T_{n0}^{cc'}(X);M)M_0^{c'}(T_{n0}^{cc'}(X))R_{0n}^{c'c}(T_{n0}^{cc'}(X))$$

$$+\sum_{n'=1}^{N}\sum_{c'\in\mathcal{D}_{n'}} \Pi(T_{nn'}^{cc'}(X);M)M_{n'}^{c'}(A_{n'}^{c'}(X_{n'}))N_{n'}^{c'}(T_{nn'}^{cc'}(X))R_{n'n}^{c'c}(T_{nn'}^{cc'}(X)),$$

and thus the global process satisfies local balance

$$\sum_{n'=0}^{N}\sum_{c'\in\mathcal{A}_{n'}} \{\Pi(X;M)Q(X,T_{nn'}^{cc'}(X);M) - \Pi(T_{nn'}^{cc'}(X);M)Q(T_{nn'}^{cc'}(X),X;M)\} = 0.$$

7.2.3 The network

Combining the descriptions of the nodes and their interaction, we obtain a queueing network of nodes in which the detailed behaviour of the node is specified in Definition 7.1, and the interaction among the nodes is specified in Definition 7.3. This network allows a Markovian description with state $x = (x_1,\ldots,x_N)$. Denote $G(x) = (G_1(x_1),\ldots,G_N(x_N))$.

Definition 7.4 (Network). The network $\mathcal{N}$ is the Markov-chain with state-space $\mathcal{S} \subseteq \{x = (x_1,\ldots,x_N) : x_n \in \mathcal{S}_n, G(x) \in \mathcal{S}_g\}$, and transition rates $q(x,x')$ from state $x = (x_1,\ldots,x_N)$ to state $x' = (x_1',\ldots,x_N')$ given by

$$q(x,x') = \sum_{c\in\mathcal{D}_n,\ c'\in\mathcal{A}_{n'}} d_n^c(x_n,x_n')N_n(G(x))R_{nn'}^{cc'}(G(x))a_{n'}^{c'}(x_{n'},x_{n'}'),$$

if $x'_n \neq x_n$, $x'_{n'} \neq x_{n'}$, and $x'_k = x_k$, for $k \neq n, n'$,

$$q(x,x') = i_n(x_n,x'_n)N_n(G(x)) + \sum_{c\in\mathcal{D}_0} M_0^c(G(x)) \sum_{c'\in\mathcal{A}_n} R_{0n}^{cc'}(G(x))a_n^{c'}(x_n,x'_n)$$

$$+ \sum_{c\in\mathcal{D}_n} d_n^c(x_n,x'_n)N_n(G(x)) \sum_{c'\in\mathcal{A}_0} R_{n0}^{cc'}(G(x)),$$

if $x'_n \neq x_n$, and $x'_k = x_k$ for $k \neq n$.

We assume that the network $\mathcal{N}$ is ergodic, and define $\pi(x)$ as the stationary probability that the network is in state x.

Arrivals and departures in the global process have been characterised via assumptions on the nominal departure rates, $M_n^c(X_n)$, and their time-reversed counterparts, $\Lambda_n^c(X_n;M)$, that are restricted to depend on the global state X_n, only. In contrast, arrivals and departures in the local processes have been characterised via assumptions on the arrival rates $\lambda_n^c(X_n)$, and their time-reversed counterparts $\mu_n^c(X_n;\lambda_n)$. This may seem somewhat inconvenient at first glance. However, arrivals to a node at local level are determined by departures from nodes at global level and subsequent routing of customers at global level. In our analysis below, we will make this relation explicit, thus characterising the relation between λ and M. Further, note that characterisation of local processes via arrival rates in the forward and time-reversed process provides a direct link with quasi-reversibility, whereas characterisation of the global process via departure rates in the forward and time-reversed processes provides a link with local balance. We may thus view our network as a network of further generalised quasi-reversible nodes linked via a process that satisfies a generalised form of local balance.

The aim of this chapter is twofold. First, we want to establish sufficient conditions on the arrival rate functions $\lambda_n^c(X_n)$, $\mu_n^c(X_n;\lambda_n)$, and the nominal departure rate functions $M_n^c(X_n)$, $\Lambda_n^c(X_n;M)$ under which the network can be decomposed, i.e. the stationary distribution $\pi(x)$ of the network can be factorised into the stationary distributions $\pi_n(x_n;\lambda_n)$ of the local processes, and the stationary distribution $\Pi(X;M)$ of the global process. Second, our aim is to investigate when the global process and the local processes are aggregations of the network, i.e., the distribution and the rates of the global process describe the evolution of the global state of the network, and the distribution and the rates of the local processes describe the evolution of the detailed state of a node in the network. Roughly said, these aggregations require that not only the stationary distribution of the network $\mathcal{N}$ can be decomposed into the stationary distributions of the local and global processes, but also the process $\mathcal{N}$ itself can be decomposed into the processes $\mathcal{L}_n(\lambda_n)$ and $\mathcal{G}(M)$.

7.3 Decomposition

This section considers the decomposition of the stationary distribution $\pi(x)$ of the network $\mathcal{N}$ into the stationary distributions of the global process and the local processes. We show that such a decomposition holds if the nominal departure rates $M_n^c(X)$ and the nominal time-reversed departure rates $\Lambda_n^c(X;M)$ of the global process equal the corresponding rates in the local processes, to be specified below. As an illustration, in Section 7.5 we consider the two models that are studied in [3] and [12]. These models fall into our class of queueing networks via specific assumptions on the form of the global state. We will show that for both models the conditions of our general result are satisfied if and only if the assumptions that are made in [3] and [12] are satisfied. In addition, we will describe pull networks [4], and derive some new decomposition results for so-called assembly networks.

The conditional probability of x_n given X_n for local process $\mathcal{L}_n(\lambda_n)$ equals $\pi_n(x_n;\lambda_n)/p_n(X_n;\lambda_n)$. Let $\tilde{M}_n^c(X_n;\lambda_n)$ denote the conditional expected class $c \in \mathcal{D}_n$ departure rate given state X_n of the local process $\mathcal{L}_n(\lambda_n)$. Then

$$\begin{aligned}\tilde{M}_n^c(X_n;\lambda_n) &= \sum_{\{x_n:G_n(x_n)=X_n\}} \frac{\pi_n(x_n;\lambda_n)}{p_n(X_n;\lambda_n)} \sum_{x_n'\in\mathcal{S}_n} d_n^c(x_n,x_n') &(7.9)\\ &= \frac{1}{p_n(X_n;\lambda_n)} \sum_{\{x_n':G_n(x_n')=D_n^c(X_n)\}} \pi_n(x_n';\lambda_n)\mu_n^c(D_n^c(X_n);\lambda_n)\\ &= \frac{p_n(D_n^c(X_n);\lambda_n)}{p_n(X_n;\lambda_n)}\mu_n^c(D_n^c(X_n);\lambda_n), &(7.10)\end{aligned}$$

where the second equality is obtained from the defintion of μ_n^c given in (7.4). Similarly, let $\tilde{\Lambda}_n^c(X_n;\lambda_n)$ denote the conditional expected class $c \in \mathcal{D}_n$ arrrival rate given state X_n of the local process $\mathcal{L}_n(\lambda_n)$. Then

$$\begin{aligned}\tilde{\Lambda}_n^c(X_n;\lambda_n) &= \sum_{\{x_n\in\mathcal{S}_n:G_n(x_n)=X_n\}} \frac{\pi_n(x_n;\lambda_n)}{p_n(X_n;\lambda_n)} \sum_{x_n'\in\mathcal{S}_n} \frac{\pi_n(x_n';\lambda_n)}{\pi_n(x_n;\lambda_n)}\lambda_n^c(G_n(x_n'))a_n^c(x_n',x_n)\\ &= \frac{p_n(D_n^c(X_n);\lambda_n)}{p_n(X_n;\lambda_n)}\lambda_n^c(D_n^c(X_n)), &(7.11)\end{aligned}$$

where the last equality is due to the restrictions on x_n' due to $a_n^c(x_n',x_n)$, i.e., $x_n' \in \{x : G_n(x) = D_n^c(X_n)\}$, and due to $a_n^c(x_n',x_n)$ being honest.

It is interesting to observe that under Assumption 7.2.1 resp. Assumption 7.2.2 we obtain flow balance under time-reversal as specified below for the local processes, resp. the global process. These observations start from the global balance equations for the local processes, for $\pi_n(x_n;\lambda_n)$ the stationary distribution of local process $\mathcal{L}_n(\lambda_n)$,

$$\pi_n(x_n;\lambda_n) \sum_{x_n'\in\mathcal{S}_n} \left(\sum_{c\in\mathcal{A}_n} \lambda_n^c(G_n(x_n))a_n^c(x_n,x_n') + \sum_{c\in\mathcal{D}_n} d_n^c(x_n,x_n') + i_n(x_n,x_n') \right)$$

$$= \sum_{x'_n \in \mathcal{S}_n} \pi_n(x'_n;\lambda_n) \left(\sum_{c \in \mathcal{A}_n} \lambda_n^c(G_n(x'_n)) a_n^c(x'_n, x_n) \right.$$

$$\left. + \sum_{c \in \mathcal{D}_n} d_n^c(x'_n, x_n) + i_n(x_n, x'_n) \right), \tag{7.12}$$

and the global balance equations for the global process, for $\Pi(X;M)$ the stationary distribution of the global process $\mathcal{G}(M)$,

$$\Pi(X;M) \sum_{n=0}^{N} \sum_{c \in \mathcal{D}_n} M_n^c(X) N_n(X)$$

$$= \sum_{n=0}^{N} \sum_{c \in \mathcal{A}_n} \sum_{n'=0}^{N} \sum_{c' \in \mathcal{D}_{n'}} \Pi(T_{nn'}^{cc'}(X);M) M_{n'}^{c'}(T_{nn'}^{cc'}(X)) N_{n'}^{c'}(T_{nn'}^{cc'}(X)) R_{n'n}^{c'c}(T_{nn'}^{cc'}(X)). \tag{7.13}$$

Summing the global balance equations (7.12) for fixed X_n over all x_n with $G_n(x_n) = X_n$, the internal transitions cancel out. The definition of $\mu_n^c(X_n;\lambda_n)$ in Assumption 7.2.1 then yields, noting that $G_n(x_n) = X_n$,

$$\sum_{\{x_n : G_n(x_n) = X_n\}} \pi_n(x_n;\lambda_n) \sum_{c \in \mathcal{A}_n} \lambda_n^c(G_n(x_n)) + \sum_{c \in \mathcal{D}_n} \sum_{\{x'_n : G_n(x'_n) = D_n^c(X_n)\}} \pi_n(x'_n;\lambda_n)$$

$$\times \mu_n^c(D_n^c(G_n(x_n));\lambda_n) = \sum_{c \in \mathcal{A}_n} \sum_{\{x'_n : G_n(x'_n) = D_n^c(X_n)\}} \pi_n(x'_n;\lambda_n)$$

$$\times \lambda_n^c(D_n^c(G_n(x_n))) + \sum_{\{x_n : G_n(x_n) = X_n\}} \pi_n(x_n;\lambda_n) \mu_n^c(G_n(x_n);\lambda_n).$$

The definition of $p_n(X_n;\lambda_n)$ now implies that for the local process $\mathcal{L}_n(\lambda_n)$ the sum of the total arrival rates and the total mean departure rates in each global state X_n does not change under time reversal:

$$\sum_{c \in \mathcal{A}_n} \lambda_n^c(X_n) + \sum_{c \in \mathcal{D}_n} \frac{p_n(D_n^c(X_n);\lambda_n)}{p_n(X_n;\lambda_n)} \mu_n^c(D_n^c(X_n);\lambda_n)$$

$$= \sum_{c \in \mathcal{D}_n} \mu_n^c(X_n;\lambda_n) + \sum_{c \in \mathcal{A}_n} \frac{p_n(D_n^c(X_n);\lambda_n)}{p_n(X_n;\lambda_n)} \lambda_n^c(D_n^c(X_n)) \tag{7.14}$$

To obtain our decomposition result, we will assume that for the global process the arrival rate to node n equals the departure rate to node n, as characterized via the time-reversed process:

$$M_n^c(X_n) = \tilde{M}_n^c(X_n;\lambda_n) \tag{7.15}$$

$$\Lambda_n^c(X_n;M) = \tilde{\Lambda}_n^c(X_n;\lambda_n). \tag{7.16}$$

Invoking (7.10), (7.15), (7.11), and (7.16) we obtain

$$\begin{aligned}\sum_{c\in\mathcal{A}_n}\lambda_n^c(G_n(x_n)) - \sum_{c\in\mathcal{A}_n}\Lambda_n^c(G_n(x_n);M)\\ = \sum_{c\in\mathcal{D}_n}\mu_n^c(G_n(X_n);\lambda_n) - \sum_{c\in\mathcal{D}_n}M_n^c(G_n(x_n)),\end{aligned} \tag{7.17}$$

i.e., *the net input due to the local and global processes equals the net output due to the local and global processes.*

A further consequence of (7.14) is that the $p_n(X_n;\lambda_n)$ can be computed recursively:

$$p_n(X_n;\lambda_n) = \frac{\sum_{c\in\mathcal{A}_n} p_n(D_n^c(X_n);\lambda_n)\lambda_n^c(D_n^c(X_n)) - \sum_{c\in\mathcal{D}_n} p_n(D_n^c(X_n);\lambda_n)\mu_n^c(D_n^c(X_n);\lambda_n)}{\sum_{c\in\mathcal{A}_n}\lambda_n^c(X_n) - \sum_{c\in\mathcal{D}_n}\mu_n^c(X_n;\lambda_n)}$$

for $\sum_{c\in\mathcal{A}_n}\lambda_n^c(X_n) \neq \sum_{c\in\mathcal{D}_n}\mu_n^c(X_n;\lambda_n)$. For $\sum_{c\in\mathcal{A}_n}\lambda_n^c(X_n) = \sum_{c\in\mathcal{D}_n}\mu_n^c(X_n;\lambda_n)$, we find

$$p_n(X_n;\lambda_n) = \frac{\sum_{c\in\mathcal{A}_n} p_n(D_n^c(X_n);\lambda_n)\lambda_n^c(D_n^c(X_n))}{\sum_{c\in\mathcal{D}_n}\tilde{M}_n^c(X_n;\lambda_n)},$$

which, for example, is the case for quasi-reversible nodes.

Invoking Assumption 7.2.2 on the nominal departure rates $\Lambda_n^c(X;M)$ in the right-hand side of the global balance equations (7.13) implies that for the global process $\mathcal{G}(M)$, the total departure rate in each state X does not change under time-reversal:

$$\begin{aligned}\sum_{c\in\mathcal{D}_0}M_0^c(X) + \sum_{n=1}^{N}\sum_{c\in\mathcal{D}_n}M_n(X)N_n(X) =\\ = \sum_{c\in\mathcal{A}_0}\Lambda_0^c(X;M) + \sum_{n=1}^{N}\sum_{c\in\mathcal{A}_n}\Lambda_n^c(X;M)N_n(X).\end{aligned} \tag{7.18}$$

We are now ready to state the main theorem of this section.

Theorem 7.3.1 *Assume that, for $n = 1,\ldots,N$, $X_n \in G_n(\mathcal{S}_n)$,*

$$\begin{aligned}M_n^c(X_n) &= \tilde{M}_n^c(X_n;\lambda_n)\\ \Lambda_n^c(X_n;M) &= \tilde{\Lambda}_n^c(X_n;\lambda_n).\end{aligned}$$

Then the stationary distribution of the network $\mathcal{N}$ is

$$\pi(x) = \Pi(G(x);M)\prod_{n=1}^{N}\frac{\pi_n(x_n;\lambda_n)}{p_n(G_n(x_n);\lambda_n)}, \quad x\in\mathcal{S}. \tag{7.19}$$

Observe that (7.15) and (7.16) place severe restrictions on the departure rates from a node in the local processes and the global process, and thus relate the sets of functions $M = (M_n^c : G_n(\mathcal{S}_n) \to \mathbb{R}_0^+; c\in\mathcal{D}_n, n = 1,\ldots,N)$ to the sets of functions $\lambda_n = (\lambda_n^c : G_n(\mathcal{S}_n) \to \mathbb{R}_0^+; c\in\mathcal{A}_n)$, $n = 1,\ldots,N$.

Proof of Theorem 7.3.1. It is sufficient to show that $\pi(x)$ solves the balance equations for the network, that read when inserting the proposed form (7.19), and dividing by $\pi(x)$:

$$\begin{aligned}
&\sum_{c\in\mathcal{D}_0} M_0^c(G(x)) + \sum_{n=1}^{N} N_n(G(x)) \sum_{x_n'} \Bigg(\sum_{c\in\mathcal{D}_n} d_n^c(x_n,x_n') + i_n(x_n,x_n') \Bigg)\\
&= \sum_{n'=1}^{N} \sum_{c'\in\mathcal{D}_{n'}} \sum_{c\in\mathcal{A}_0} \frac{\Pi(T_{0n'}^{cc'}(G(x));M)}{\Pi(G(x);M)} N_{n'}^{c'}(T_{0n'}^{cc'}(G(x))) R_{n'0}^{c'c}(T_{0n'}^{cc'}(G(x)))\\
&\quad\times \frac{p_{n'}(G_{n'}(x_{n'});\lambda_{n'})}{p_{n'}(A_{n'}^{c'}(G_{n'}(x_{n'}));\lambda_{n'})} \Bigg(\sum_{x_{n'}'} \frac{\pi_{n'}(x_{n'}';\lambda_{n'})}{\pi_{n'}(x_{n'};\lambda_{n'})} d_{n'}^{c'}(x_{n'}',x_{n'}) \Bigg)\\
&\quad+ \sum_{n=1}^{N} \sum_{c\in\mathcal{A}_n} \sum_{x_n'} \frac{\pi_n(x_n';\lambda_n)}{\pi_n(x_n;\lambda_n)} a_n^c(x_n',x_n) \frac{p_n(G_n(x_n);\lambda_n)}{p_n(D_n^c(G_n(x_n));\lambda_n)} \Bigg(\sum_{c'\in\mathcal{D}_0} \frac{\Pi(T_{n0}^{cc'}(G(x)))}{\Pi(G(x))}\\
&\quad\times M_0^{c'}(T_{n0}^{cc'}(G(x))) R_{0n}^{c'c}(T_{n0}^{cc'}(G(x))) + \sum_{n'=1}^{N} \sum_{c'\in\mathcal{D}_{n'}} \frac{\Pi(T_{nn'}^{cc'}(G(x));M)}{\Pi(G(x);M)} N_{n'}^{c'}(T_{nn'}^{cc'}(G(x)))\\
&\quad\times R_{n'n}^{c'c}(T_{nn'}^{cc'}(G(x)))\ \frac{p_{n'}(G_{n'}(x_{n'});\lambda_n)}{p_{n'}(A_{n'}^{c'}(G_{n'}(x_{n'}));\lambda_n)} \Bigg(\sum_{x_{n'}'} \frac{\pi_{n'}(x_{n'}';\lambda_{n'})}{\pi_{n'}(x_{n'};\lambda_{n'})} d_{n'}^{c'}(x_{n'}',x_{n'}) \Bigg)\Bigg)\\
&\quad+ \sum_{n=1}^{N} \sum_{x_n'} \frac{\pi_n(x_n')}{\pi_n(x_n)} N_n(G(x)) i_n(x_n',x_n).
\end{aligned}$$

Invoking (7.4), (7.10), and (7.15), and (7.7), the first term on the right hand side equals $\sum_{c\in\mathcal{A}_0} \Lambda_0^c(G(x);M)$. Invoking (7.4), (7.10), (7.15), (7.8), (7.16) and (7.11), the second and third term in the right hand side equal:

$$\sum_{n=1}^{N} \sum_{c\in\mathcal{A}_n} \sum_{x_n'} \frac{\pi_n(x_n';\lambda_n)}{\pi_n(x_n;\lambda_n)} a_n^c(x_n',x_n) \lambda_n^c(D_n^c(X_n);M) N_n(X).$$

Inserting these expressions in the right hand side, and invoking global balance for the nodes (7.12), implies that it is sufficient to show that

$$\begin{aligned}
&\sum_{c\in\mathcal{D}_0} M_0^c(G(x)) + \sum_{n=1}^{N} \sum_{c\in\mathcal{D}_n} \sum_{x_n'} \frac{\pi_n(x_n')}{\pi_n(x_n)} d_n^c(x_n',x_n) N_n(G_n(x_n))\\
&= \sum_{n=1}^{N} N_n(G(x)) \sum_{c\in\mathcal{A}_n} \lambda_n^c(G_n(x_n)) + \sum_{c\in\mathcal{A}_0} \Lambda_0^c(G(x);M) \qquad (7.20)
\end{aligned}$$

Inserting (7.17) into (7.18) yields (7.20), which completes the proof. □

The decomposition of Theorem 7.3.1 does not establish a complete decomposition of the nodes and the global process, in the sense that the state of the nodes and the global state of the network are independent. Equation (7.19) states that the detailed states of the nodes are independent, conditioned on the global state of the nodes:

$$\frac{\pi(x)}{\Pi(X;M)} = \prod_{n=1}^{N} \frac{\pi_n(x_n;\lambda_n)}{p_n(X_n;\lambda_n)}.$$

The proof of Theorem 7.3.1 relies heavily on (7.14) but does not require additional properties of $p_n(X_n;\lambda_n)$. An immediate generalisation of Theorem 7.3.1 is obtained replacing $p_n(X_n;\lambda_n)$ by any function satisfying (7.14) .

Theorem 7.3.2 *Let $f_n : \mathcal{S}_n \to \mathbb{R}_0^+$ be a function satisfying*

$$\sum_{c\in\mathcal{A}_n} \lambda_n^c(X_n) + \sum_{c\in\mathcal{D}_n} \frac{f_n(D_n^c(X_n);\lambda_n)}{f_n(X_n;\lambda_n)} \mu_n^c(D_n^c(X_n);\lambda_n)$$

$$= \sum_{c\in\mathcal{D}_n} \mu_n^c(X_n;\lambda_n) + \sum_{c\in\mathcal{A}_n} \frac{f_n(D_n^c(X_n);\lambda_n)}{f_n(X_n;\lambda_n)} \lambda_n^c(D_n^c(X_n)), \tag{7.21}$$

and assume that the following conditions are satisfied:

$$M_n^c(X_n) = \frac{f_n(D_n^c(X_n);\lambda_n)}{f_n(X_n;\lambda_n)} \mu_n^c(D_n^c(X_n);\lambda_n), \tag{7.22}$$

$$\Lambda_n^c(X_n;M) = \frac{f_n(D_n^c(X_n);\lambda_n)}{f_n(X_n;\lambda_n)} \lambda_n^c(D_n^c(X_n)). \tag{7.23}$$

Then the stationary distribution $\pi(x)$ of the network $\mathcal{N}$ is

$$\pi(x) = C^{-1}\Pi(G(x);M)\prod_{n=1}^{N} \frac{\pi_n(x_n;\lambda_n)}{f_n(G_n(x_n);\lambda_n)}, \quad x \in \mathcal{S}, \tag{7.24}$$

with

$$C = \sum_{x\in\mathcal{S}} \Pi(G(x);M)\prod_{n=1}^{N} \frac{\pi_n(x_n;\lambda_n)}{f_n(G_n(x_n);\lambda_n)}.$$

For generalised quasi-reversible nodes conditions (7.22), (7.23) are satisfied with $f_n(X_n;\lambda_n) = 1$. In this case, a complete decomposition can be obtained from Theorem 7.3.2.

As a consequence of Theorem 7.3.2, we can simplify the formula for the stationary distribution in case the local processes are extended quasi-reversible: then $f_n(X_n) = 1$ satisfies condition (7.21), and the following Corollary follows immediately from Theorem 7.3.2.

Corollary 7.3.3 *Assume that the local processes $\mathcal{L}_n(\lambda_n)$ are generalised quasi-reversible, say with arrival rates $\lambda_n^c(X_n) = \hat{\lambda}_n^c$ and time-reversed arrival rates*

$\mu_n^c(X_n;\lambda_n) = \hat{\mu}_n^c$. *Let* $M_n^c(X_n)$ *be given by*

$$M_n^c(X_n) = \begin{cases} \hat{\mu}_n^c & \text{if } D_n^c(X_n) \in G_n(\mathcal{S}_n), \\ 0 & \text{otherwise.} \end{cases}$$

If for all X_n *with* $D_n^c(X_n) \in G_n(\mathcal{S}_n)$, $\Lambda_n^c(X_n;M) = \lambda_n^c$, *then the stationary distribution of the network* $\mathcal{N}$ *is given by*

$$\pi(x) = C^{-1}\Pi(G(x))\prod_{n=1}^{N}\pi_n(x_n;\lambda_n), \quad x \in \mathcal{S},$$

with

$$C = \sum_{x\in\mathcal{S}}\Pi(G(x))\prod_{n=1}^{N}\pi_n(x_n;\lambda_n).$$

Theorems 7.3.1, 7.3.2 and Corollary 7.3.3 require the values for the arrival rates $\lambda_n = (\lambda_n^c : G_n(\mathcal{S}_n) \to \mathbb{R}_0^+; c \in \mathcal{A}_0)$ that relate the local processes and the global process. These arrival rates are the solution of the fixed point problem consisting of the equations.

Corollary 7.3.4 (Fixed point equations for arrival rates: decomposition) *The arrival rates* $\lambda_n = (\lambda_n^c : G_n(\mathcal{S}_n) \to \mathbb{R}_0^+; c \in \mathcal{A}_0)$ *are a solution of the fixed point equations:*

$$\begin{array}{ll}
(7.12) & \pi_n = \pi_n(\lambda_n) \\
(7.4) & \mu_n^c = \mu_n^c(\pi_n,\lambda_n) \\
(7.10) & \tilde{M}_n^c = \tilde{M}_n^c(\pi_n,\mu_n,\lambda_n) \\
(7.15) & M = \tilde{M} \\
(7.13) & \Pi = \Pi(M) \\
(7.7) & \Lambda_n^c = \Lambda_n^c(M,\Pi) \\
(7.16) & \tilde{\Lambda} = \Lambda \\
(7.11) & \lambda = \lambda(\tilde{\Lambda})
\end{array}$$

These equations may be solved using the following algorithm:

Step i: For $n = 1,\ldots,N$ *initialize with a starting value for* $\hat{\lambda}_n$ *for* λ_n.
Step ii: Use (7.12), (7.4), (7.10) to obtain $\tilde{M}_n^c$.
Step iii: Use (7.15), (7.13), (7.7) to obtain Λ_n^c.
Step iv: If $\tilde{\Lambda}_n^c = \Lambda_n^c$ *for* $c \in \mathcal{D}_n$, $n = 1,\ldots,N$ *then stop, and* λ_n *is obtained, else use (7.11) to let*

$$\lambda_n^c(X_n) = \frac{p_n(A_n^c(X_n);\lambda_n)}{p_n(X_n;\lambda_n)}\Lambda_n^c(A_n^c(X_n);M).$$

and go to Step ii.

Notice that existence of a fixed point is an implicit assumption that we made for the results of Theorems 7.3.1, 7.3.2 to be valid.

The arrival rates and time-reversed arrival rates of our network depend on the state x through the global state $G(x)$ only. For the network to satisfy Assumption 7.2.1 additional assumptions on the global state are required. Section 7.5 provides examples of networks that have this structure.

7.4 Aggregation

This section considers aggregation of the nodes in our network. We first show that under the conditions of Theorem 7.3.1, the global process is the aggregation of the network with respect to the global state, that is, for the analysis of the global network the detailed behaviour of the nodes is not required. We then investigate under which conditions the local processes are the aggregation of the network with respect to the detailed state of a single node, that is, for the analysis of the detailed behaviour of a single node the detailed behaviour of the *other* nodes is not required. It appears that this requires some extra restrictions on the arrival rates: the local arrival rates should equal the global arrival rates. Our generalisation results in an aggregation algorithm that generalises the method developed by Marie in [20].

The following definition is adapted from Brandwajn [9].

Definition 7.5 (Aggregation). Consider two Markov chains $\mathcal{M}_1$ and $\mathcal{M}_2$ with state spaces S_1 and S_2, transition rates $q_1(y_1,y_1')$, $y_1,y_1' \in S_1$, and $q_2(y_2,y_2')$, $y_2,y_2' \in S_2$, and stationary distributions $\pi_1(y_1), y_1 \in S_1$, and $\pi_2(y_2), y_2 \in S_2$. The Markov chain $\mathcal{M}_2$ is said to be the aggregation of $\mathcal{M}_1$ with respect to a function $h : S_1 \to S_2$ if the following two conditions are satisfied:

$$\pi_2(y_2) = \sum_{\{y_1 \in S_1 : h(y_1) = y_2\}} \pi(y_1), \quad y_2 \in S_2, \tag{7.25}$$

$$\pi_2(y_2)q_2(y_2,y_2') = \sum_{\{y_1,y_1' \in S_1 : h(y_1) = y_2, h(y_1') = y_2'\}} \pi_1(y_1)q(y_1,y_1'), \quad y_2,y_2' \in S_2. \tag{7.26}$$

The definition of aggregation requires both the equilibrium distribution and the probability flows to match. Boucherie [3] refers to this form of aggregation as first order equivalence. The intuition for Theorem 7.4.1 is encapsulated in (7.15), (7.16) of Theorem 7.3.1: $M_n^c(X_n) = \tilde{M}_n^c(X_n;\lambda_n)$, $\Lambda_n^c(X_n;M) = \tilde{\Lambda}_n^c(X_n;\lambda_n)$. These equations state that for the global process the arrival rate to node n equals the departure rate to node n, as characterized via the time-reversed process, which expresses conservation of probability flow.

Theorem 7.4.1 (Aggregation with respect to the global state function) *Assume that, for $n = 1,\ldots,N$, $X_n \in G_n(\mathcal{S}_n)$,*

$$\begin{aligned} M_n^c(X_n) &= \tilde{M}_n^c(X_n;\lambda_n) \\ \Lambda_n^c(X_n;M) &= \tilde{\Lambda}_n^c(X_n;\lambda_n). \end{aligned}$$

Then the global process $\mathcal{G}(M)$ is the aggregation of the network $\mathcal{N}$ with respect to the global state function $G : \mathcal{S} \to \mathcal{S}_g$.

Proof. Condition (7.25) is almost immediate:

$$\sum_{\{x:G(x)=X\}} \pi(x) = \Pi(X) \sum_{\{x:G(x)=X\}} \prod_{n=1}^{N} \frac{\pi_n(x_n;\lambda_n)}{p_n(X_n;\lambda_n)} = \Pi(X), \quad X \in \mathcal{S}_g.$$

For condition (7.26), we first consider a transition from global state X to $T^{cc'}_{nn'}(X)$ with $n, n' \neq 0$, $c \in \mathcal{D}_n$, and $c' \in \mathcal{A}_{n'}$. The aggregate probability flow for this transition is

$$\sum_{\substack{\{x,x' : G(x) = X, \\ G(x') = T^{cc'}_{nn'}(G(x))\}}} \Pi(X;M) \prod_{i=1}^{N} \frac{\pi_i(x_i;\lambda_i)}{p_i(X_i;\lambda_i)} d^c_n(x_n,x'_n) N_n(X) R^{cc'}_{nn'}(X) a^{c'}_{n'}(x_{n'},x'_{n'})$$

$$= \Pi(X;M) N_n(X) R^{cc'}_{nn'}(X) \sum_{\substack{\{x_n,x'_n : G_n(x_n) = X_n, \\ G_n(x'_n) = D^c_n(X_n)\}}} \frac{\pi_n(x_n;\lambda_n)}{p_n(X_n;\lambda_n)} d^c_n(x_n,x'_n) \Pi(X) N_n(X) R^{cc'}_{nn'}(X) M^c_n(X_n),$$

which is the corresponding probability flow in the global process $\mathcal{G}(M)$. For transitions from state X to state $T^{cc'}_{0n'}(X)$ and state $T^{cc'}_{n0}(X)$, condition (7.26) is proved analogously. □

Let us now study conditions for the local processes to be the aggregation of the network with respect to the detailed state of a node. The multiplication factor $N_n(X)$ in the transition rates for the network is not incorporated in the local processes, so that we must set $N_n(X) = N_n(X_n)$. We will restrict the network to

$$N_n(X) = 1, \quad n = 1,\dots,N, \quad X \in \mathcal{S}_g.$$

For aggregation with respect to the nodes, we need additional conditions. To this end, observe that Theorem 7.3.1 has been obtained under the condition that the departure and time-reversed departure rates of the local processes equal the corresponding rates in the global processes. Intuitively, for the local processes to be the aggregation of the network with respect to the nodes, it is also required that the local arrival rates equal the corresponding rates in the global process. Let us first specify the arrival rates in the global process that will be used in the formulation of our aggregation result.

Let $\tilde{\lambda}^c_n(X_n)$ denote the mean class $c \in \mathcal{A}_n$ arrival rate at node n in state X_n, $n = 1,\dots,N$, of the global process $\mathcal{G}(M)$. Then

$$\tilde{\lambda}_n^c(X_n) = \sum_{Y:Y_n=X_n} \frac{\Pi(Y;M)}{P_n(X_n;M)} \left(\sum_{c'\in\mathcal{D}_0} M_0^{c'}(X) R_{0n}^{c'c}(X) + \sum_{n'=1}^{N} \sum_{c'\in\mathcal{D}_{n'}} M_{n'}^{c'}(X_{n'}) R_{n'n}^{c'c}(X) \right) \tag{7.27}$$

$$= \frac{1}{P_n(X_n;M)} \sum_{Y:Y_n=A_n^c(X_n)} \left(\sum_{c'\in\mathcal{D}_0} \Pi(T_{n0}^{cc'}(Y)) M_0^{c'}(T_{n0}^{cc'}(Y)) R_{0n}^{c'c}(T_{n0}^{cc'}(Y)) \right.$$

$$\left. + \sum_{n'=1}^{N} \sum_{c'\in\mathcal{D}_{n'}} \Pi(T_{nn'}^{cc'}(Y)) M_{n'}^{c'}(A_{n'}^{c'}(Y_{n'})) R_{n'n}^{c'c}(T_{nn'}^{cc'}(Y)) \right)$$

$$= \frac{P_n(A_n^c(X_n);M)}{P_n(X_n;M)} \Lambda_n^c(A_n^c(X_n);M), \tag{7.28}$$

where the term

$$\sum_{c'\in\mathcal{D}_0} M_0^{c'}(X) R_{0n}^{c'c}(X) + \sum_{n'=1}^{N} \sum_{c'\in\mathcal{D}_{n'}} M_{n'}^{c'}(X_{n'}) R_{n'n}^{c'c}(X)$$

in the first line (7.27) is the class c arrival rate at node n in state X of the global process, and the last equality follows from the definitions of $\Lambda_n^c(X_n;M)$ and $P_n(X_n;M)$.

Under the conditions (7.15), (7.16) of Theorem 7.3.1 the local class $c \in \mathcal{A}_n$ arrival rate $\lambda_n^c(X_n)$ is related to $\Lambda_n^c(X_n;M)$ by

$$\lambda_n^c(X_n) = \frac{p_n(A_n^c(X_n);\lambda_n)}{p_n(X_n;\lambda_n)} \Lambda_n^c(A_n^c(X_n);M).$$

The following theorem shows that if this rate equals the corresponding rate $\tilde{\lambda}_n^c(X_n)$ as specified in (7.28) for the global process, the local processes are the aggregation of the network with respect to the nodes. Note that this further implies that the aggregate probability $p_n(X_n;\lambda_n)$ that the local process is in state X_n equals the corresponding probability $P_n(X_n;M)$ for the global process.

Theorem 7.4.2 (Aggregation with respect to the detailed state of the nodes) *Assume that $N_n(X) = 1$ for all n and X, and that, for $n = 1,\dots,N$, $X_n \in G_n(\mathcal{S}_n)$,*

$$\begin{aligned} M_n^c(X_n) &= \tilde{M}_n^c(X_n;\lambda_n) \\ \Lambda_n^c(X_n;M) &= \tilde{\Lambda}_n^c(X_n;\lambda_n). \end{aligned}$$

Further assume that for $n = 1,\dots,N$, $X_n \in G_n(\mathcal{S}_n)$,

$$\tilde{\lambda}_n^c(X_n) = \lambda_n^c(X_n). \tag{7.29}$$

Then, for $n = 1,\dots,N$, $X_n \in G_n(\mathcal{S}_n)$,

$$P_n(X_n;M) = p_n(X_n;\lambda_n) \tag{7.30}$$

and the local process $\mathcal{L}_n$ is the aggregation of $\mathcal{N}$ with respect to the aggregation function $h(x) = x_n$.

Proof. First observe that condition (7.29) implies that

$$\frac{p_n(A_n^c(X_n);\lambda_n)}{p_n(X_n;\lambda_n)} = \frac{P_n(A_n^c(X_n);M)}{P_n(X_n;M)}.$$

Since both $p_n(\cdot)$ and $P_n(\cdot)$ are probabilities over $G_n(\mathcal{S}_n)$, it must be that (7.30) is satisfied.

The aggregate probability that the state of node n in the network equals x_n is given by

$$\begin{aligned}\sum_{\{y:y_n=x_n\}} \pi(y) &= \sum_{\{y:y_n=x_n\}} \Pi(G(y);M)\prod_{i=1}^{N}\frac{\pi_i(y_i;\lambda_i)}{p_i(G_i(y_i);\lambda_i)}\\ &= \sum_{\{Y:Y_n=G_n(x_n)\}} \Pi(Y;M)\frac{\pi_n(x_n;\lambda_n)}{p_n(G_n(x_n);\lambda_n)}\\ &= P_n(G_n(x_n);M)\frac{\pi_n(x_n;\lambda_n)}{p_n(G_n(x_n);\lambda_n)} = \pi_n(x_n;\lambda_n),\end{aligned}$$

the corresponding probability in the local process. Hence, condition (7.25) is satisfied.

It remains to prove that condition (7.26) is satisfied. For internal transitions, note that the probability flow of an internal transition of node n from state x_n to x_n' in the network is given by

$$\sum_{\{y:y_n=x_n\}}\sum_{\{y':y_n'=x_n'\}} \pi(y)i_n(y_n,y_n') = \pi_n(x_n;\lambda_n)i_n(x_n,x_n').$$

For departure transitions, condition (7.26) is proved similarly. Let us now consider a class $c \in \mathcal{A}_n$ arrival transition from state x_n to state x_n'. The probability flow of this transition in the network is given by

$$\begin{aligned}&\sum_{\{y:y_n=x_n\}}\Bigg(\sum_{c'\in\mathcal{D}_0}\sum_{\substack{\{y':\ y_n'=x_n'\\ G(y')=T_{0n}^{c'c}(G(y))\}}} \pi(y)M_0^{c'}(G(y))R_{0n}^{c'c}(G(y))a_n^c(y_n,y_n')\\ &+\sum_{n'=1}^{N}\sum_{c'\in\mathcal{D}_{n'}}\sum_{\substack{\{y':\ y_n'=x_n'\\ G(y')=T_{n'n}^{c'c}(G(y))\}}} \pi(y)d_{n'}^{c'}(y_{n'},y_{n'}')R_{n'n}^{c'c}(G(y))a_n^c(y_n,y_n')\Bigg)\\ &= \pi_n(x_n)a_n^c(x_n,x_n')\sum_{\{Y:G_n(y_n)=G_n(x_n)\}} \Pi(Y;M)\frac{1}{p_n(Y_n;\lambda_n)}\end{aligned}$$

$$\left(\sum_{c'\in\mathcal{D}_0} M_0^{c'}(Y)R_{0n}^{c'c}(Y)+\sum_{n'=1}^{N}\sum_{c'\in\mathcal{D}_{n'}} M_{n'}^{c'}(Y_{n'})R_{n'n}^{c'c}(Y)\right)$$
$$=\pi_n(x_n;\lambda_n)\lambda_n(G_n(x_n))a_n^c(x_n,x_n'),$$

which is the corresponding rate in the local process. Note that the last equality is obtained using (7.27) and (7.30). □

Note that the conditions of Theorem 7.4.2 include those of Theorem 7.4.1. Thus under the conditions of Theorem 7.4.2 both aggregations hold. Note also that the decomposition (7.19) still holds: the stationary distribution of the network thus may be factorised such that the local processes are aggregations of the network with respect to the nodes, and the global process is the aggregation of the network with respect to the global state.

Under the conditions of Theorem 7.4.2, the arrival rates $\lambda_n=(\lambda_n^c : G_n(\mathcal{S}_n)\to\mathbb{R}_0^+; c\in\mathcal{A}_0)$ are a solution of a set of fixed point equations that comprises those of Corollary 7.3.4 and in addition (7.29) and (7.30). To simplify this set of equations, note that (7.30) implies that both (7.29): $\lambda=\tilde{\lambda}$, and (7.16): $\tilde{\Lambda}=\Lambda$ are satisfied. We have the following result.

Corollary 7.4.3 (Fixed point equations for arrival rates: aggregation) *Under the conditions of Theorem 7.4.2, the arrival rates $\lambda_n=(\lambda_n^c : G_n(\mathcal{S}_n)\to\mathbb{R}_0^+; c\in\mathcal{A}_0)$ are a solution of the fixed point equations:*

$$\begin{array}{ll}
(7.12) & \pi_n=\pi_n(\lambda_n)\\
(7.4) & \mu_n^c=\mu_n^c(\pi_n,\lambda_n)\\
(7.10) & \tilde{M}_n^c=\tilde{M}_n^c(\pi_n,\mu_n,\lambda_n)\\
(7.15) & M=\tilde{M}\\
(7.13) & \Pi=\Pi(M)\\
(7.28) & \tilde{\lambda}_n^c=\tilde{\lambda}_n^c(\Lambda_n^c)\\
(7.30) & P_n=p_n\\
 & \lambda_n^c(X_n)=\frac{P_n(A_n^c(X_n);M)}{P_n(X_n;M)}\Lambda_n^c(A_n^c(X_n);M).
\end{array}$$

These equations may be solved using the following algorithm:

Step i: For $n=1,\dots,N$ initialize with a starting value for $\hat{\lambda}_n$ for λ_n.
Step ii: Use (7.12), (7.4), (7.10) to obtain $\tilde{M}_n^c$.
Step iii: Use (7.15), (7.13), (7.28) to obtain λ_n^c.
Step iv: If $P_n(X_n)=p_n(X_n)$ for $c\in\mathcal{D}_n$, $X_n\in G_n(\mathcal{S}_n)$, $n=1,\dots,N$ then stop, and λ_n is obtained, else let

$$\lambda_n^c(X_n)=\frac{P_n(A_n^c(X_n);M)}{P_n(X_n;M)}\Lambda_n^c(A_n^c(X_n);M)$$

and go to Step ii.

Remark 7.2 (Marie's decomposition and aggregation method). The algorithm of Corollary 7.4.3 requires Assumptions 7.2.1 and 7.2.2. Observe, however, that the algorithm can also be evaluated if these assumptions do not hold by replacing formula (7.10) for the mean local departure rate by (7.9), and formula (7.28) for the mean global arrival rates by (7.27). This gives an approximation algorithm that extends Marie's method [20] to include state-dependent routing, general global states and to global processes that do not satisfy local balance. □

7.5 Examples

This section provides some examples to illustrate the results of Sections 7.3 and 7.4. The first three examples relate our results to known cases from the literature that have motivated the results of this paper. Section 7.5.1 describes a network of quasi-reversible nodes linked via state-dependent routing as studied in [3]. Section 7.5.2 describes biased local balance and a network with negative customers and signals as studied in [12]. The third example in Section 7.5.3 is concerned with pull networks as studied in [4] for which the partial balance equations are different from the standard equations for Jackson type networks. Finally, Section 7.5.4 provides a novel example of assembly networks. We obtain novel product form results and novel decomposition results.

7.5.1 Quasi-reversible nodes linked via state-dependent routing

Consider a network of N interacting nodes containing customers of a single class, say $\mathcal{A}_n = \mathcal{D}_n = \{1\}$ for all $n = 0,\dots,N$. Let the global state X_n of node $n = 1,\dots,N$ represent the total number of customers in node n. Let $A_n^1(X_n) = X_n + 1$, $D_n^1(X_n) = X_n - 1$, i.e. an arriving customer increases the number of customers by one, and a departing customer decreases the number of customers by one. For simplicity, we also assume that $G_n(\mathcal{S}_n) = \{0,\dots,M\}$, where M may represent infinity. This assumption, however, is not essential for the results below.

In (7.14) we have shown that for the local process $\mathcal{L}_n(\lambda_n)$ the sum of the total arrival rates and the total mean departure rates in each global state X_n does not change under time reversal. This implies for a network containing only a single class of customers that the local time-reversed arrival rates equal the time-forward arrival rates:

$$\mu_n^1(X_n;\lambda_n) = \lambda_n^1(X_n) \quad n = 1,\dots,N. \tag{7.31}$$

To see this, first note that for $X_n = 0$, the result follows since $p_n(-1) = 0$. Now suppose $\mu_n^1(X_n;\lambda_n) = \lambda_n^1(X_n)$ for $X_n < M$. Then, again by (7.14), $\mu_n^1(X_n+1;\lambda_n) = \lambda_n^1(X_n+1)$, since $p_n(X_n+1) > 0$ by the ergodicity of the local processes.

Equation (7.31) states that the outside of the nodes in the local process should satisfy local balance (with possibly state-dependent arrival rates). In the following lemma we show that this property is equivalent to quasi-reversibility (i.e., with constant arrival rates).

Lemma 7.5.1 *Assume that $\mathcal{A}_n = \mathcal{D}_n = \{1\}$ for all $n = 0,\dots,N$. Let the global state X_n of node $n = 1,\dots,N$ represent the total number of customers in node n. Let $A_n^1(X_n) = X_n + 1$, $D_n^1(X_n) = X_n - 1$. Then $\mu_n^1(X_n;\lambda_n) = \lambda_n^1(X_n)$ if and only if node n is quasi-reversible when the arrival rate equals one.*

Proof. Suppose node n is quasi-reversible with arrival rate one, and $\pi_n(x_n;1)$ is its stationary distribution. By substitution in the balance equations, we obtain that

$$\pi_n(x_n;1)\prod_{y=0}^{G_n(x_n)-1}\lambda_n^1(y) \tag{7.32}$$

is the stationary distribution of node n with arrival rate $\lambda_n^1(X_n)$, and that $\mu_n(X_n;\lambda_n) = \lambda_n(X_n)$. Similarly, if $\mu_n(X_n;\lambda_n) = \lambda_n(X_n)$ and node n has stationary distribution $\pi_n(x_n;\lambda_n)$, then

$$\pi_n(x_n;\lambda_n)\left(\prod_{y=0}^{G_n(x_n)-1}\lambda_n^1(y)\right)^{-1}$$

is the stationary distribution of node n with arrival rate 1, and $\mu_n(X_n;1) = 1$. □

Let us now consider the implications of a single class with global state representing the number of customers in the global process. By Lemma 7.5.1 we immediately see that

$$\Lambda_n^c(X_n;M) = M_n^c(X_n),$$

since for (generalised) quasi-reversible nodes we may invoke Theorem 7.3.2 with $f_n(X_n;\lambda_n) = 1$, or Corollary 7.3.3. The global process thus must satisfy local balance. The following lemma shows that the choice of the departure rates $M_n^c(X_n)$ does not effect the local balance of the global process: local balance of the global process is a property that is only determined by the coupling of the nodes, and not by the nodes themselves.

Lemma 7.5.2 *Assume that $\mathcal{A}_n = \mathcal{D}_n = \{1\}$ for all $n = 0,\dots,N$. Let the global state X_n of node $n = 1,\dots,N$ represent the total number of customers in node n. Let $A_n^1(X_n) = X_n + 1$, $D_n^1(X_n) = X_n - 1$. Then $M_n^1(X_n) = \Lambda_n^1(X_n;M)$ if and only if the global process satisfies local balance when $M_n(X_n) = 1$.*

Proof. Suppose the global process satisfies local balance with $M_n(X_n) = 1$, and let $\Pi(X;\mathbf{1})$ denote the stationary distribution when $M_n(X_n) = 1$. Then it is readily verified by substitution in the balance equations for the global process that

$$\Pi(X;\mathbf{1})\prod_{n=1}^{N}\left(\prod_{y=1}^{X_n}M_n(y)\right)^{-1} \tag{7.33}$$

is the stationary distribution for the global process with departure rates $M_n(X_n)$, and that $\Lambda_n(X_n;M) = M_n(X_n)$. Similarly, if $\Pi(X;M)$ is the stationary distribution of the global process with departure rates $M_n(X_n)$, and $\Lambda_n(X_n;M) = M_n(X_n)$, then

$$\Pi(X;M)\prod_{n=1}^{N}\left(\prod_{y=1}^{X_n}M_n(y)\right)$$

is the stationary distribution for the global process with departure rates equal to one, and satisfies local balance for the global process. □

We summarize the above results in the following theorem, that states the conditions on the nodes and the local processes of [3]. We want to stress that the results presented above need all conditions stated here. Theorem 7.5.3 generally will not hold for multiclass queueing networks, networks with batch movements, or networks with negative customers.

Theorem 7.5.3 *Assume that $\mathcal{A}_n = \mathcal{D}_n = \{1\}$ for all $n = 0,\dots,N$. Let the global state X_n of node $n = 1,\dots,N$ represent the total number of customers in node n. Let $A_n^1(X_n) = X_n + 1$, $D_n^1(X_n) = X_n - 1$. The conditions of Theorems 7.3.1 and 7.3.2, and of Corollary 7.3.3 are satisfied if and only if the nodes are quasi-reversible with arrival rate one, and the global process satisfies local balance with departure rates one.*

If $\lambda_n^1(X_n) = \mu_n^1(X_n;\lambda_n)$, then any function f_n satisfies (7.21). Hence, Theorem 7.3.2 allows the global process to be analysed by arbitrary departure rate functions. From (7.32) and (7.33) we find that the stationary distribution in Theorem 7.3.2 takes the form

$$C\Pi(G(x);\mathbf{1})\prod_{n=1}^{N}\left(\prod_{y=1}^{G_n(x_n)}\lambda_n^1(y-1)\frac{f_n(y-1)}{f_n(y)}\right)^{-1}\frac{\pi_n(x_n;1)}{f_n(G_n(x_n))}\prod_{y=0}^{G_n(x_n)-1}\lambda_n^1(y)$$

$$= C\prod_{n=1}^{N}f_n(0)\Pi(G(x);\mathbf{1})\prod_{n=1}^{N}\prod_{n=1}^{N}\pi_n(x_n;1),$$

in correspondence with Corollary 7.3.3.

7.5.2 Biased local balance

For the global process, we have assumed in Assumption 7.2.2 that the nominal departure rate of class c customers from node n in state X of the stationary time-reversed process of $\mathcal{G}(M)$ depends on the global state X_n only, i.e.,

$$\Lambda_n^c(X_n;M)N_n(X) = \sum_{c'\in\mathcal{D}_0}\frac{\Pi(T_{n0}^{cc'}(X);M)}{\Pi(X;M)}M_0^{c'}(T_{n0}^{cc'}(X))R_{0n}^{c'c}(T_{n0}^{cc'}(X))$$

$$+\sum_{n'=1}^{N}\sum_{c'\in\mathcal{D}_{n'}}\frac{\Pi(T_{nn'}^{cc'}(X);M)}{\Pi(X;M)}M_{n'}^{c'}(A_{n'}^{c'}(X_{n'}))N_{n'}^{c'}(T_{nn'}^{cc'}(X))R_{n'n}^{c'c}(T_{nn'}^{cc'}(X)).$$

We have shown that, if $\mathcal{A}_n=\mathcal{D}_n$, and $\Lambda_n^c(X_n;M)=M_n^c(X_n)$, $X_n\in G_n(\mathcal{S}_n)$, $c\in\mathcal{A}_n$, $n=1,\ldots,N$, then this assumption implies that the global process satisfies local balance (where, for notational convenience, $N_0^c(X)=1$ for all c,X)

$$\begin{aligned}&M_n^c(X_n)N_n(X)\Pi(X;M)\\&=\sum_{n'=0}^{N}\sum_{c'\in\mathcal{D}_{n'}}\Pi(T_{nn'}^{cc'}(X);M)M_{n'}^{c'}(A_{n'}^{c'}(X_{n'}))N_{n'}^{c'}(T_{nn'}^{cc'}(X))R_{n'n}^{c'c}(T_{nn'}^{cc'}(X)),\end{aligned}$$

otherwise we do not have equality. Following Chao and Miyazawa [12] we introduce *biased local balance,* and say that $\Pi(X;M)$ satisfies biased local balance with bias $\gamma_n^c(X;M)$ if

$$\begin{aligned}&(M_n^c(X_n)N_n(X)+\Gamma_n^c(X;M))\,\Pi(X;M)\\&=\sum_{n'=0}^{N}\sum_{c'\in\mathcal{D}_{n'}}\Pi(T_{nn'}^{cc'}(X);M)M_{n'}^{c'}(A_{n'}^{c'}(X_{n'}))N_{n'}^{c'}(T_{nn'}^{cc'}(X))R_{n'n}^{c'c}(T_{nn'}^{cc'}(X)),\end{aligned} \tag{7.34}$$

Our definition of biased local balance is closely related to the concept of biased local balance, introduced by Chao and Miyazawa [12]. However, in [12] the bias is required to be constant, and thus the existence of the bias imposes conditions on the global process. By allowing the bias to be state-dependent, the bias can be defined for every global process.

Note that global balance implies that

$$\sum_{n=0}^{N}\sum_{c\in\mathcal{A}_n\cup\mathcal{D}_n}\Gamma_n^c(X;M)=0,\quad X\in\mathcal{S}_g. \tag{7.35}$$

Further note that Assumption 7.2.2 implies that

$$\Gamma_n^c(X;M)=(\Lambda_n^c(X_n;M)-M_n^c(X_n))\,N_n(X), \tag{7.36}$$

i.e., we have a strict condition on the state dependence of the bias.

We now define the bias of the local process as the difference in arrival and departure rates to a node. For the local processes $\mathcal{L}_n(\lambda_n)$, we call $\gamma_n^c(x_n;\lambda_n)$ the *bias of node n with respect to the outside and* c, if for all $x_n\in\mathcal{S}_n$

$$\pi_n(x_n;\lambda_n)\,(\lambda_n^c(G_n(x_n))+\gamma_n^c(x_n;\lambda_n))=\sum_{x_n'}\pi_n(x_n';\lambda_n)d_n^c(x_n',x_n). \tag{7.37}$$

Similar to the bias of the global process, the bias indicates the unbalance in local balance equations: if $\gamma_n^c(x_n;\lambda_n)=0$, equation (7.37) corresponds to node n being locally balanced with respect to its outside and type u, and thus, if in addition λ_n^c is a constant function, (7.37) corresponds to node n being quasi-reversible in the definition of [18]. When $\gamma_n^c(x_n;\lambda_n)$ is constant, but not necessarily zero, and λ_n^c is a constant function, (7.37) states that node n is quasi-reversible according to the generalised definition of [11]. Again, allowing the bias to be state-dependent, it can be defined for every node n, without requiring conditions on this node. Assumption 7.2.1 implies that

$$\gamma_n^c(x_n;\lambda_n) = \mu_n^c(G_n(x_n);\lambda_n) - \lambda_n^c(G_n(x_n)) \tag{7.38}$$

From our assumptions, invoking (7.10), (7.15), (7.11), and (7.16) we have obtained (7.17), that may be rewritten as

$$N_n(X)\gamma_n^c(X_n;\lambda_n) = -\Gamma_n^c(X;M)$$

i.e., the bias of the local process equals the bias of the global process. Our results of Section 7.3 thus show that if the bias of the nodes is suitably compensated by the bias of the global process, the network allows a decomposition of the stationary distribution.

Chao and Miyazawa [12] introduced the concept of biased local balance to extend the definition of quasi-reversibility allowing the input and output rate of customers at the nodes to differ from each other. The model of [12] has no global state for the nodes, say $G_n = 0$. Routing then is necessarily state-independent, and the multiplication factors $N_n(X)$ may be omitted, i.e., we may set $N_n(X) = 1$. Removing the global state also implies removing the state-dependence of the arrival and departure rates. The following theorem summarizes the product form result of [12].

Theorem 7.5.4 *Assume that $X_n = 0$ for all n. Then the conditions of Theorems 7.3.1, 7.3.2 and Corollary 7.3.3 are satisified if and only if each node is generalised quasi-reversible, say with $\hat{\lambda}_n^c$ and $\hat{\mu}_n^c$, and the following traffic equations hold:*

$$\lambda_n^c = \sum_{n'=0}^{N} \sum_{c' \in \mathcal{D}'_n} \mu_n^c R_{n'n}^{c'c} \tag{7.39}$$

7.5.3 A pull network

In a Jackson network a transition is initiated by the service of a customer at a node, and subsequently this customer is routed to its destination. This behaviour is sometimes referred to as push network: a customer is pushed from one queue to the next queue. We now consider a pull network in which a transition is initiated by the destination node that pulls a customer from another node.

Consider a network of N interacting nodes containing customers of a single class, say $\mathcal{A}_n = \mathcal{D}_n = \{1\}$ for all $n = 0,\ldots,N$. Let the global state X_n of node $n = 1,\ldots,N$ represent the total number of customers in node n. Let $A_n^1(X_n) = X_n - 1$, $D_n^1(X_n) = X_n + 1$. A departure from node n increases the number of customers in node n by one, and with probability $R_{nn'}^{11}(X)$ decreases the number of customers in node n' by one:

node n thus pulls a customer with probability $R^{11}_{nn'}(X)$ from node n'. For simplicity, we also assume that $G_n(\mathcal{S}_n) = \{0,\dots,M\}$, where M may represent infinity. The following results are easily proved in the same way as in Section 7.5.1.

First, we may show that $\mu^1_n(X_n;\lambda_n) = \lambda^1_n(X_n)$ for all $n = 1,\dots,N$, and $\mu^1_n(X_n;\lambda_n) = \lambda^1_n(X_n)$ if and only if node n is quasi-reversible when the arrival rate equals one. Furthermore, we have that $M^1_n(X_n) = \Lambda^1_n(X_n;M)$ if and only if the global process satisfies local balance when $M_n(X_n) = 1$. Summarizing, we have the following result.

Theorem 7.5.5 *Assume that $\mathcal{A}_n = \mathcal{D}_n = \{1\}$ for all $n = 0,\dots,N$. Let the global state X_n of node $n = 1,\dots,N$ represent the total number of customers in node n. Let $A^1_n(X_n) = X_n - 1$, $D^1_n(X_n) = X_n + 1$. The conditions of Theorem 7.3.1 are satisfied if and only if the nodes are quasi-reversible with arrival rate one, and the global process satisfies local balance with departure rates one.*

Thus, the seemingly distinct formulations of the local balance equations for push and pull networks that are described in [4] are a consequence of the same notion of local balance.

7.5.4 An assembly network

Consider a simple assembly network consisting of three nodes. Node 1 and node 2 each represent a subnetwork, on which we make no other assumption than that they produce units at nominal rate one. The units produced by node 1 are referred to as class 1 units; the units that are produced by node 2 as class 2 units. Both nodes send their units to node 3, where a class one and a class two unit are assembled into a class 3 unit. Assembly takes an exponentially distributed time with mean $\beta^{-1} < 1$, and clearly requires that both a class 1 and a class 2 unit are present at node 3.

We assume the following control mechanism in the network. If there are no class 1 units in node 3, node 2 is slowed down by a factor $\phi < 1$. Similarly, if no class 2 units are present in node 3, node 1 is slowed down by the same factor ϕ. This control mechanism thus tries to save production costs by producing less units when these units do not directly lead to output. We will show that for a specific choice of ϕ the network has a product from solution, and the time-reversed class 3 arrival rate is constant.

Let us first consider the local processes. For node 1 and node 2 we need no arrival transitions. We will omit the λ_n, $n = 1,2$, from the notation. The stationary distributions π_1 and π_2 of the local processes for node $n = 1,2$ thus are the unique distributions satisfying

$$\pi_n(x_n) \sum_{x'_n \in \mathcal{S}_n} \left(i_n(x_n,x'_n) + d^n_n(x_n,x'_n)\right) = \sum_{x'_n \in \mathcal{S}_n} \pi_n(x'_n)\left(i_n(x'_n,x_n) + d^n_n(x'_n,x_n)\right).$$

By the assumption that nodes 1 and 2 produce units at nominal rate one, we have

$$\sum_{x'_n \in \mathcal{S}_n} \frac{\pi_n(x'_n)}{\pi_n(x_n)} d_n^n(x'_n, x_n) = 1.$$

Therefore, no global state for node 1 and 2 is required (note that the routing is fixed, and the control mechanism is only influenced by node 3). As the global state for nodes 1, 2 is not required, we may set $X_1 = X_2 = 0$, and, hence, $p_1(0) = 1$, $p_2(0) = 1$.

The state of node 3 is described by $x_3 = (u_1, u_2)$, with u_n denoting the number of class n units in node 3. Since upon arrival of a class $n = 1,2$ unit, the number of class n units is increased by one, arrival transitions are given by

$$a_3^n(x_n, x_n + e_n) = 1,$$

with e_n denoting the n-th unit vector of dimension 2. Departure transitions take place at rate β, as long as there are both type 1 and a type 2 units present in node 3. As a class 3 departure reduces the number of class 1 and class 2 units by one, the departure transitions are thus given by

$$d_3^3((u_1,u_2),(u_1-1,u_2-1)) = \beta, \quad u_1, u_2 > 0.$$

Internal transitions do not occur, as the service times of node 3 are exponential. To model the desired control mechanism, we define the global state of node 3 equal to the detailed state. Note that this is allowed by the exponential service times, and the unique changes of the state at arrival and service transitions. Then $p_3 = \pi_3$ and the functions $A_3^c(X_3)$ and $D_3^c(X_3)$ for $c = 1,2,3$ are given by

$$A_3^c(X_3) = \begin{cases} X_3 + e_c & \text{for } c = 1,2, \\ X_3 + e_1 + e_2 & \text{for } c = 3, \end{cases}$$

$$D_3^c(X_3) = \begin{cases} X_3 - e_c & \text{for } c = 1,2, \\ X_3 - e_1 - e_2 & \text{for } c = 3. \end{cases}$$

To define a local process for node 3, we need an initial guess for the arrival rates of class 1 and class 2 units. An obvious choice is the following.

$$\lambda_3^1((u_1,u_2)) = \begin{cases} \phi & \text{for } u_2 = 0 \\ 1 & \text{otherwise} \end{cases} \tag{7.40}$$

$$\lambda_3^2((u_1,u_2)) = \begin{cases} \phi & \text{for } u_1 = 0 \\ 1 & \text{otherwise} \end{cases} \tag{7.41}$$

The stationary distribution of the resulting local process $\mathcal{L}_3(\lambda_3)$ is provided in the following lemma for a specific choice of ϕ.

Lemma 7.5.6 *Let $\lambda_3 = (\lambda_3^1, \lambda_3^2)$ be given by (7.40) and (7.41). For $\phi = \frac{1}{2}\beta\alpha^2$, with*

$$\alpha = -\frac{1}{2} + \frac{1}{2\beta}\sqrt{\beta^2 + 8\beta}, \tag{7.42}$$

the stationary distribution π_3 of the local process $\mathcal{L}_3(\lambda_3)$ is given by

$$\pi_3((u_1,u_2);\lambda_3) = (1-\alpha)^2\alpha^{u_1+u_2}. \tag{7.43}$$

Under these conditions, the time-reversed class 3 arrival rate $\mu_3^3((u_1,u_2);\lambda_3)$ is constant and equal to $\beta\alpha^2$.

Proof. As (7.43) sums to one, it is sufficient to prove that (7.43) satisfies the balance equations. For $u_1,u_2 > 0$, these equations are given by

$$\begin{aligned}&\pi((u_1,u_2);\lambda_3)(2+\beta)\\ &\quad = \pi((u_1-1,u_2);\lambda_3)+\pi((u_1,u_2-1);\lambda_3)+\beta\pi((u_1+1,u_2+1);\lambda_3).\end{aligned}$$

Substitution of (7.43) and dividing by $(1-\alpha^2)\alpha^{u_1+u_2-1}$ results in

$$\alpha(2+\beta) = 2+\beta\alpha^3.$$

This implies that either $\alpha = 1$, or

$$\beta\alpha^2+\beta\alpha-2 = 0. \tag{7.44}$$

As α, as given by (7.42) solves this equation, the proposed form for π_3 satisfies the balance equations for $u_1,u_2 > 0$. For $u_1 = u_2 = 0$, the balance equations are easily seen to be satisfied for $\phi = \frac{1}{2}\beta\alpha^2$. For $u_2 = 0$ and $u_1 > 0$, the balance equations are given by

$$\pi_3((u_1,0);\lambda_3)(\phi+1) = \pi((u_1+1,1);\lambda_3)\beta+\pi((u_1-1,0);\lambda_3)\phi.$$

Substituting (7.43) and dividing by $(1-\alpha)^2\alpha^{u_1-1}$, we have

$$\alpha(\phi+1) = \alpha^3\beta+\phi.$$

As $\phi = \frac{1}{2}\beta\alpha^2$, this equation is equivalent to (7.44) and thus satisfied by the form of α. As the model is symmetric in u_1 and u_2, the first statement is proved.

By definition, the time-reversed arrival rate is given by

$$\mu_n((u_1,u_2);\lambda_3) = \frac{\pi_3((u_1+1,u_2+1);\lambda_3)}{\pi_3((u_1,u_2);\lambda_3)}\beta.$$

The second statement of the Theorem now follows from (7.43). □

Let us now consider the network and the global process. The routing functions are obviously given by $R_{13}^{11} = 1$, $R_{23}^{22} = 1$ and $R_{30}^{33} = 1$. Furthermore, the control mechanism is incorporated in the model by

$$N_1((u_1,u_2)) = \begin{cases}\phi & \text{for } u_2 = 0\\ 1 & \text{otherwise}\end{cases}$$

$$N_2((u_1,u_2)) = \begin{cases} \phi \text{ for } u_1 = 0 \\ 1 \text{ otherwise,} \end{cases}$$

and $N_3((u_1,u_2)) = 1$. Note that we have defined no global state for node 1 and 2, and thus the global state of the network is given by the global state of node 3. According to Theorem 7.3.1, the departure rate function are given by $M_1^1(0) = 1$, $M_2^2(0) = 1$, and using Lemma 7.5.6 we find

$$M_3^3((u_1,u_2)) = \frac{\pi_3((u_1-1,u_2-1);\lambda_3)}{\pi_3((u_1,u_2);\lambda_3)}\alpha^2 = \begin{cases} \beta \text{ for } u_1,u_2 > 0 \\ 0 \text{ otherwise} \end{cases}$$

Constructing the rates of the global process by Definition 7.3, and using the definition of the time-reversed departure rates of the global process, we obtain the following lemma.

Lemma 7.5.7 *The global process equals the local process for node* 3 *and satisfies Assumption 7.2.2 with*

$$\Lambda_3^1((u_1,u_2);M) = \begin{cases} \phi\alpha^{-1} \textit{ for } u_2 = 0, \\ \alpha^{-1} \quad \textit{otherwise,} \end{cases}$$

$$\Lambda_3^2((u_1,u_2);M) = \begin{cases} \phi\alpha^{-1} \textit{ for } u_1 = 0, \\ \alpha^{-1} \quad \textit{otherwise.} \end{cases}$$

According to Theorem 7.3.1, the class 1 arrival rate of the local process for node 3 corresponding with $\Lambda_3^1((u_1,u_2);M)$ should be equal to

$$\frac{p_3((u_1+1,u_2);\lambda_3)}{p_3((u_1,u_2);\lambda_3)}\Lambda_3^1((u_1,u_2);M) = \begin{cases} \phi \text{ for } u_2 = 0, \\ 1 \text{ otherwise.} \end{cases}$$

Similarly, the local class 2 arrival rate should be equal to ϕ for $u_1 = 0$ and equal to 1 otherwise. Hence, our initial guess for these local arrival rates was correct, and by Theorem 7.3.1 and Lemma 7.5.6, we have the following result.

Theorem 7.5.8 *The stationary distribution of the assembly network is of product-form:* $\pi_1(x_1)\pi_2(x_2)\pi_3(x_3)$. *The time-reversed class* 3 *arrival rate of the network is constant and equals* $\beta\alpha^2$. □

References

1. S. Balsamo and G. Iazeolla, An extension of Norton's theorem for queueing networks, IEEE Trans. on Software Eng. 8 (1982) 298-305.
2. Baskett, F., Chandy, K. M., Muntz, R. R. and Palacios, F. G. (1975) Open, closed and mixed networks of queues with different classes of customers. *J. A. C. M.* **Vol. 22** 248-260.
3. Boucherie, R.J. (1998) Norton's equivalent for queueing networks comprised of quasi-reversible components linked by state-dependent routing. *Performance Evaluation*, 32, 83-99.

4. Boucherie, R.J., Chao, X. and Miyazawa, M. (2003) Arrival first queueing networks with applications in kanban production systems. *Performance Evaluation*, 51, 83-102.
5. Boucherie, R.J. and Van Dijk, N.M. (1991) Product forms for queueing networks with state-dependent multiple job transitions. *Advances in Applied Probability* **23**, 152-187.
6. Boucherie, R.J. and Van Dijk, N.M. (1993) A generalization of Norton's theorem for queueing networks. *Queueing Systems* **13**, 251-289.
7. Boucherie, R.J. and van Dijk, N.M. (1994) Local balance in queueing networks with positive and negative customers. *Annals of Operations Research*, 48, 463-492.
8. A. Brandt, On Norton's theorem for multi-class queueing networks of quasi-reversible nodes, Preprint Nr. 256, Humboldt-Universität zu Berlin, Sektion Mathematik (1990).
9. Brandwajn, A. (1985) Equivalence and decomposition in queueing systems – a unified approach. *Performance Evaluation*, 5, 175-186.
10. K. M. Chandy, U. Herzog, and L. Woo, Parametric analysis of queueing networks, IBM J. Res. Develop. 19 (1975) 36-42.
11. Chao, X. and Miyazawa, M. (1996) A probabilistic decomposition approach to quasi-reversibility and its applications in coupling of queues. Technical report, New Jersey Institute of Technology and Science University Tokyo.
12. Chao, X. and Miyazawa, M. (1998) On quasi-reversibility and local balance: An alternative derivation of the product-form results. *Operations Research*, 46, 927-933.
13. van Dijk, N.M. (1991) Product forms for queueing networks with limited clusters. *Research Report, Vrije Universiteit, Amsterdam* http://hdl.handle.net/1871/12242.
14. van Dijk, N.M. (1993) *Queueing networks and product forms: a systems approach.* John Wiley & Sons.
15. Gelenbe, E. (1989) Random neural networks with negative and positive signals and product form solution. *Neural Computation* **1**, 502-510.
16. Hsiao, M.-T. T. and Lazar A.A. (1989) An Extension to Norton's Equivalent. *Queueing Systems*, 5, 401-412.
17. Jackson, J. R. (1963) Jobshop-like queueing systems. *Management Science* **10** 131-142.
18. Kelly, F.P. (1979) *Reversibility and stochastic networks.* John Wiley & Sons.
19. P. S. Kritzinger, S. Van Wyk and A. E. Krzesinski, A generalisation of Norton's theorem for multiclass queueing networks, Performance Evaluation 2 (1982) 98-107.
20. Marie, R.A. (1979) An approximate analytical method for general queueing networks. *IEEE Transactions on Software Engineering*, SE-5, 530-538.
21. Pellaumail, J. (1979) Formule de produit et décomposition de résaux de files d'attente. *Annales de l'Institut Henri Poincaré*, 15, 261–286.
22. Towsley, D. (1980) Product form and local balance in queueing networks. *Journal of the ACM*, 27, 323–337.
23. H. Vantilborgh, Exact aggregation in exponential queueing networks, J. A. C. M. 25 (1978) 620-629.
24. J. Walrand, A note on Norton's theorem for queueing networks, J. Appl. Prob. 20 (1983) 442-444.
25. J. Walrand and P. Varaiya, Interconnections of Markov chains and quasi-reversible queueing networks, Stoch. Proc. Appl. 10 (1980) 209-219.
26. Whittle, P. (1967) Nonlinear migration processes. *Bull. Inst. Int. Statist.* **42**, 642-647.
27. Whittle, P. (1968) Equilibrium distributions for an open migration process. *J. Appl. Prob.* **5**, 557-571.
28. P. Whittle, Systems in stochastic equilibrium, Wiley (1986).

Chapter 8
Stochastic Comparison of Queueing Networks

Ryszard Szekli

Abstract We recall classical queueing networks and their stochastic monotonicity properties as a special case of a general stochastic ordering theory for Markov processes. As a consequence of stochastic monotonicity we present stochastic bounds in transient and stationary conditions for the queue length processes, and some dependence and ordering properties for sojourn times in networks. We overview properties of throughputs in networks in connection with stochastic orderings. Finally we concentrate on dependence orderings for queueing networks with a special attention on the role of routing as a parameter influencing correlation structures in networks. Some connections to the problem of speed of convergence to stationarity via spectral gaps are pointed out.

8.1 Introduction

Classical network theory. A. K. Erlang developed the basic foundations of teletraffic theory long before probability theory was popularized or even well developed. He established many of principal results which we still use today. The 1920's were basically devoted to the application of Erlang's results (Molina [64], Thornton Fry [29]). Felix Pollaczek [70] did further pioneering work, followed by Khintchine [42] and Palm [67]. It was until the mid 1930's, when Feller introduced the birth-death process, that queueing was recognized by the world of mathematics as an object of serious interest. During and following World War II this theory played an important role in the development of the new field of operations research, which seemed to hold so much promise in the post war years. The frontiers of this research proceeded into far reaches of deep and complex mathematics. Not all of these developments proved to be useful. The fact that one of the few tools available for analyzing the

Ryszard Szekli
University of Wrocław, Mathematical Institute, pl. Grunwaldzki 2/4, 50-384 Wrocław, Poland
e-mail: Ryszard.Szekli@math.uni.wroc.pl

R.J. Boucherie and N.M. van Dijk (eds.), *Queueing Networks: A Fundamental Approach*,
International Series in Operations Research & Management Science 154,
DOI 10.1007/978-1-4419-6472-4_8, © Springer Science+Business Media, LLC 2011

performance of computer network systems is queueing theory largely stimulated development of it. Important contributions in 1950's and 60's are among others due to V. E. Benes , D. G. Kendall , D. R. Lindley , S. Karlin and J. L. McGregor, R. M. Loynes , J. F. C. Kingman, L. Takacs, R. Syski, N. U. Prabhu and J. W. Cohen. The literature grew from ”solutions looking for a problem” rather than from ”problems looking for a solution”, which remains true in some sense nowadays. The practical world of queues abounds with problems that cannot be solved elegantly but which must be analyzed. The literature on queues abounds with ”exact solutions”, ”exact bounds”, simulation models, etc., with almost everything but little common sense methods of ”engineering judgment”. It is very often that engineers resort to using formulas which they know they are using incorrectly, or run to the computer even if they need only to know something to within a factor of two. There is a need for approximations, bounds, heuristic reasoning and crude estimates in modelling. The present chapter is an overview of methods based on stochastic ordering which are useful in obtaining comparisons and bounds. Early other efforts following the line of finding estimates are formulated in Newell [66], and Gross, Harris [32] where fluid and diffusion approximations were introduced. The theory of weak convergence has been a strong impetus for a systematic development of limit theorems for queueing processes (Whitt [99]). Point processes have played an important role in the description of input and output processes. Palm measures and Palm-martingale calculus (see e.g. Baccelli and Bremaud [4]) still play active role in stochastic network modelling not only because they are indispensable as a tool for solving stability questions but also because the Palm theory proved to be an appropriate tool to formalize arguments while proving dependence properties of queueing characteristics and showing bounds on them, as it will be presented in this chapter. In more recent literature, martingale calculus influences modelling of fluid flow queues but this is another topic not touched in this chapter.

Traffic processes. Traffic is a key ingredient of queueing systems. While traditional analytical models of traffic were often devised and selected for the analytical tractability they induced in the corresponding queueing systems, this selection criterion is largely absent from recent (internet) traffic models. In particular, queueing systems with offered traffic consisting of autoregressive type processes or self-similar processes are difficult to solve analytically. Consequently, these are only used to derive simulation models. On the other hand some fluid models are analytically tractable, but only subject to considerable restrictions. Thus the most significant traffic research problem is to solve analytically induced systems, or in the absence of a satisfactory solution, to devise approximate traffic models which lead to analytically tractable systems. Comparison of complex systems with simpler ones or finding simple bounds on sojourn times or throughput seems to be important. We shall stress this point in the present chapter.

Traditional traffic models (renewal, Markov, autoregressive, fluid) have served well in advancing traffic engineering and understanding performance issues, primarily in traditional telephony. The advent of modern high speed communications networks results in a highly heterogeneous traffic mix. The inherent burstiness of several important services makes more noticeable some serious modelling inadequacies of traditional models, particularly in regard to temporal dependence. This situation has brought about renewed interest in traffic modelling and has driven the development of new models. Statisticians are now aware that ignoring long range dependence can have drastic consequences for many statistical methods. However traffic engineers and network managers will only be convinced of the practical relevance of fractal traffic models by direct arguments, concerning the impact of fractal properties on network performance. Thus fractal traffic (stochastic modelling, statistical inference) has been a new task for researchers. While non-fractal models have inherently short-range dependence, it is known that adding parameters can lead to models with approximate fractal features. A judicious choice of a traffic model could lead to tractable models capable of approximating their intractable counterparts (and may work for some performance aspects). Therefore there is still a need to study traditional classical queueing network models. It is worth mentioning that long range dependence properties of traffic processes can be basically different when viewed under the continuous time stationary regime versus the Palm stationary regime therefore it is once again important to use the Palm theory.

Classical Networks. The networks described by Kelly [40], by Jackson [36] and by Gordon-Newell [66] are classical. These networks still remain in the range of interest of many researchers as basic tractable models, because of many interesting features such as product form, insensitivity, Poisson flows: Burke's [9], product form for sojourn times (see Serfozo [76] where Palm measures, stochastic intensities and time reversal are utilized). Large scale networks are interesting from a topological point of view. Internet seen as a random graph has its vertex distribution following a power law. This is a surprising fact stimulating researches to use random graph theory, spectral graph theory and other methods to build new models, however researching classical models with ”large” parameters remains to be important. One of the most important features of classical networks is a widespread property of being in some sense stochastically monotone. Various monotonicity and stochastic ordering results for queues are scattered in many books and very numerous papers in the existing literature, see for example parts of books by Baccelli and Bremaud [4], Chen and Yao [14], Glasserman and Yao [30], Last and Brandt [49], Müller and Stoyan [65], Ridder [73], Shaked and Shanthikumar [77], Szekli [88], Van Doorn [91] among others.

The number of articles on various aspects of stochastic ordering for queueing systems is so large that a task of over-viewing them does not seem to be a reasonable one. Therefore, this chapter concentrates only on results which are essentially for multi-node networks, excluding pure single systems results. Even with this restriction this text is certainly not complete in any sense. Formal definitions of classical networks models are recalled in order to unify notation. Networks with breakdowns are less known and the product formula for them is rather new.

We shall use notation marked with tilde for open networks in order to avoid misunderstanding in formulations where both open and closed networks appear. It is useful especially for routeing matrices, since there are some subtle differences between them for open and closed systems.

It is very often that for simple models even elementary questions are not easy to answer. In order to illustrate this point consider a simple example of an open queueing network which is the Simon–Foley [87] network of single server queues, see Fig. 8.1. A customer traversing path $(1,2,3)$ can be overtaken by customers proceeding directly to node 3 when departing from node 1. This is one of the reasons why the traffic structure in a network can be very complicated and not easy to analyze. Simon and Foley [87] proved that the vector (ξ_1, ξ_2, ξ_3) of the successive sojourn times for a customer traversing path $(1,2,3)$ has positively correlated components ξ_1 and ξ_3.

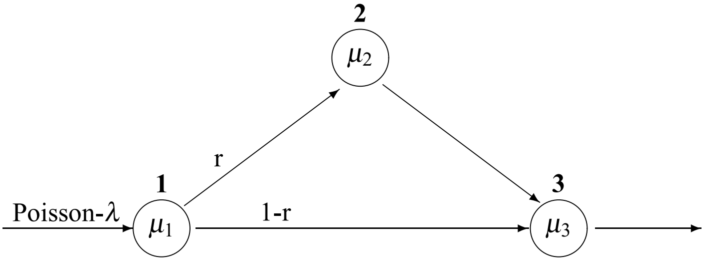


Fig. 8.1: The Simon–Foley network with overtaking due to the network topology

While the Simon–Foley network provides us with an example where overtaking is due to the topological structure of the network, an early example of Burke [11] (see Fig. 8.2) shows that overtaking due to the internal node structure prevents sojourn times on a linear path from independence as well: a three–station path $(1,2,3)$ with a multiserver node 2 $(m_2 > 1)$ has dependent components ξ_1 and ξ_3.

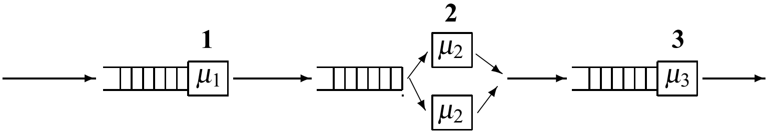


Fig. 8.2: The tandem network with overtaking due to the internal node structure

The question whether on the three–station path of the Simon–Foley network the complete sojourn time vector (ξ_1, ξ_2, ξ_3) is associated remains unanswered. We shall give some related results on sojourn times later in this chapter, also for closed networks. Before doing this we shall recall a general description of classical queueing networks, and shall discuss in a detail the topic of stochastic monotonicity of networks which is a basic property connected with stochastic comparison of networks.

8.1.1 Jackson networks

Consider a **Jackson network** which consists of J numbered nodes, denoted by $\mathrm{J} = \{1,\ldots,J\}$. Station $j \in \mathrm{J}$, is a single server queue with infinite waiting room under FCFS (First Come First Served) regime. Customers in the network are indistinguishable. There is an external Poisson arrival stream with intensity $\lambda > 0$ and arriving customers are sent to node j with probability $\tilde{r}_{0j}$, $\sum_{j=1}^{J} \tilde{r}_{0j} = r \le 1$. The quantity $\tilde{r}_{00} := 1 - r$ is then the rejection probability with that customers immediately leave the network again. Customers arriving at node j from the outside or from other nodes request a service time which is exponentially distributed with mean 1. Service at node j is provided with intensity $\mu_j(n_j) > 0$ ($\mu_j(0) := 0$), where n_j is the number of customers at node j including the one being served. All service times and arrival processes are assumed to be independent.

A customer departing from node i immediately proceeds to node i with probability $\tilde{r}_{ij} \ge 0$ or departs from the network with probability $\tilde{r}_{i0}$. The routing is independent of the past of the system given the momentary node where the customer is. Let $\mathrm{J}_0 := \mathrm{J} \cup \{0\}$. We assume that $\tilde{R} := (\tilde{r}_{ij}, i, j \in \mathrm{J}_0)$ is irreducible.

Let $\tilde{X}_j(t)$ be the number of customers present at node j at time $t \ge 0$. Then $\tilde{X}(t) = (\tilde{X}_1(t),\ldots,\tilde{X}_J(t))$ is the joint queue length vector at time instant $t \ge 0$ and $\tilde{\mathbf{X}} := (\tilde{X}(t), t \ge 0)$ is the joint queue length process with state space $(\mathbb{E}, \prec) := (\mathbb{N}^J, \le^J)$ (where $\le^J$ denotes the standard coordinate-wise ordering, $\mathbb{N} = \{0,1,2,\ldots\}$).

The following theorem is classical (Jackson [36]).

Theorem 8.1.1 *Under the above assumptions the queueing process $\tilde{\mathbf{X}}$ is a Markov process with the infinitesimal operator $Q^{\tilde{X}} = (q^{\tilde{X}}(x,y) : x,y \in E)$ given by*

$$q^{\tilde{X}}(n_1,\ldots,n_i,\ldots,n_J; n_1,\ldots,n_i+1,\ldots,n_J) = \lambda \tilde{r}_{0i}$$

and for $n_i > 0$

$$q^{\tilde{X}}(n_1,\ldots,n_i,\ldots,n_J; n_1,\ldots,n_i-1,\ldots,n_J) = \mu_i(n_i)\tilde{r}_{i0},$$

$$q^{\tilde{X}}(n_1,\ldots,n_i,\ldots,n_j,\ldots,n_J; n_1,\ldots,n_i-1,\ldots,n_j+1,\ldots,n_J) = \mu_i(n_i)\tilde{r}_{ij}\,.$$

Furthermore

$$q^{\tilde{X}}(x,x) = -\sum_{y \in E \setminus \{x\}} q^{\tilde{X}}(x,y) \text{ and } q^{\tilde{X}}(x,y) = 0 \text{ otherwise.}$$

The parameters of a Jackson network are: the arrival intensity λ, the routing matrix $\tilde{R}$ (with its routing vector $\tilde{\eta}$), the vector of service rates $\mu = (\mu_1(\cdot),\ldots,\mu_J(\cdot))$, and the number of nodes J. We shall use $(\lambda, \tilde{R}/\mu/J)$ to denote such a Jackson network.

8.1.2 Gordon-Newell networks

By a **Gordon-Newell** network we mean a closed network with $N \geq 1$ customers cycling. The routing of the customers in this network is Markovian, governed by an irreducible stochastic matrix $R = (r_{ij}, 1 \leq i,j \leq J)$. The Gordon-Newell network process $\mathbf{X}$, denoting the numbers of customers at nodes, with state space $\mathbb{E}_N = \{\mathbf{n} = (n_1,\ldots,n_J) : n_j \in \{0,1,\ldots\}, j = 1,\ldots,J, n_1 + \ldots + n_J = N\}$ is a generalized migration process with the following transition rates:

$$q^{\mathbf{X}}(\mathbf{n}, \mathbf{n} - \mathbf{e}_i + \mathbf{e}_j) = r_{ij}\mu_i(n_i), \qquad n_i \geq 1,$$

and $q^{\mathbf{X}}(\mathbf{n}, \mathbf{n}') = 0$ for all other states, where $\mathbf{e}_j$ is the j-th base vector in $\mathbb{R}^J$.

We assume that every node can be reached from any other node in a finite number of steps with positive probability. This ensures that the set of routing (traffic) equations

$$\eta_j = \sum_{i=1}^{J} \eta_i r_{ij}, \qquad j = 1,\ldots,J, \tag{8.1}$$

has a unique probability solution which we denote by $\eta = (\eta_j : j = 1,\ldots,J)$.

If at node $j \in \{1,\ldots,J\}$, n_j customers are present (including the one in service, if any) the service rate is $\mu_j(n_j) \geq 0$; we set $\mu_j(0) = 0$. Service and routing processes are independent.

Let $\mathbf{X} = (X(t) : t \geq 0)$ denote the vector process recording the joint queue lengths in the network at time t. For $t \in \mathbb{R}_+$, $X(t) = (X_1(t),\ldots,X_J(t))$ reads: at time t there are $X_j(t)$ customers present at node j, either in service or waiting. The assumptions put on the system imply that $\mathbf{X}$ is a strong Markov process with infinitesimal operator $Q^{\mathbf{X}} = (q^{\mathbf{X}}(x,y) : x,y \in \mathbb{E}_N)$.

The parameters of a Gordon-Newell network are: the routing matrix R, the vector of service rates $\mu = (\mu_1(\cdot),\ldots,\mu_J(\cdot))$, the number of nodes J, and the number of customers N. We shall use $(R/\mu/J+N)$ to denote such a network.

8.1.3 Ergodicity of classical networks

For Jackson networks, by the *product formula* for stationary distribution we mean the next formula appearing in the following theorem.

Theorem 8.1.2 *The unique invariant and limiting distribution $\tilde{\pi}^J$ of the Jackson network state process $\tilde{\mathbf{X}}$ is given by*

$$\tilde{\pi}^J(n_1,\ldots,n_J) = K(J)^{-1} \prod_{j=1}^{J} \prod_{k=1}^{n_j} \frac{\tilde{\eta}_j}{\mu_j(k)}, \quad (n_1,\ldots,n_J) \in \mathbb{N}^J \tag{8.2}$$

with the normalization constant $K(J) = \prod_{j=1}^{J}\left(1+\sum_{n=1}^{\infty}\prod_{k=1}^{n}\frac{\tilde{\eta}_j}{\mu_j(k)}\right)$, *and with* $\tilde{\eta} = (\tilde{\eta}_0,\dots,\tilde{\eta}_J)$, *the unique solution of the routing (or traffic) equation of the network (with* $\tilde{\eta}_0 = \lambda$*):*

$$\tilde{\eta}_j = \tilde{r}_{0j}\lambda + \sum_{i=1}^{J}\tilde{\eta}_i\tilde{r}_{ij} \quad , \; j \in \mathrm{J}. \tag{8.3}$$

We have therefore that $\tilde{\pi}^J(n_1,\dots,n_J) = \prod_{j=1}^{J}\tilde{\pi}_j^J(n_j)$, for the marginal distributions

$$\tilde{\pi}_j^J(n) = \tilde{\pi}_j^J(0)\prod_{k=1}^{n}\frac{\tilde{\eta}_j}{\mu_j(k)},$$

for $n \geq 1$, and $\tilde{\pi}_j^J(0) = \left(1+\sum_{n=1}^{\infty}\prod_{k=1}^{n}\frac{\tilde{\eta}_j}{\mu_j(k)}\right)^{-1}$, $j = 1,\dots,J$.

$\tilde{\eta}$ is usually not a stochastic vector and we define the unique stochastic solution of (8.3) for $j \in \mathrm{J}_0$, by

$$\xi = (\xi_j : j = 0,1,\dots,J). \tag{8.4}$$

Regarding ergodicity of closed networks the following theorem is classical (Gordon, Newell [31]).

Theorem 8.1.3 *The process* $\mathbf{X}$ *is ergodic and its unique steady–state and limiting distribution is given by*

$$\pi^{(N,J)}(n) = G(N,J)^{-1}\prod_{j=1}^{J}\prod_{k=1}^{n_j}\frac{\eta_j}{\mu_j(k)}, \tag{8.5}$$

for $n \in \mathbb{E}_N$, *(for products with upper limit* $n_j = 0$ *we set value 1) where*

$$G(N,J) = \sum_{n_1+\dots+n_J=N}\prod_{j=1}^{J}\prod_{k=1}^{n_j}\frac{\eta_j}{\mu_j(k)}$$

is the norming constant.

Let us define independent random variables Y_j, $j = 1,\dots,J$ such that

$$\Pr(Y_j = 0) = \left(1+\sum_{n=1}^{\infty}\prod_{k=1}^{n}\frac{\eta_j}{\mu_j(k)}\right)^{-1}, \; \Pr(Y_j = n) = \Pr(Y_j = 0)\prod_{k=1}^{n}\frac{\eta_j}{\mu_j(k)}.$$

Note that we have then

$$\begin{aligned}\pi^{(N,J)}(n) &= \prod_{j=1}^{J}\Pr(Y_j = n_j)/\Pr(Y_1+\dots+Y_J = N)\\ &= \Pr(Y_1 = n_1,\dots,Y_J = n_J \mid Y_1+\dots+Y_J = N).\end{aligned}$$

Therefore the stationary distribution in a closed network can be interpreted as a conditional distribution of an open network, given the number of customers present.

A natural measure of performance for a network in stationary conditions is, for each j, $E(\mu_j(X_j(t)))$. It is well known that

$$E(\mu_j(X_j(t))) = \eta_j \Pr(Y_1 + \ldots + Y_J = N-1)/\Pr(Y_1 + \ldots + Y_J = N)$$

$$= \eta_j G(N-1,J)/G(N,J).$$

Therefore $E(\mu_j(X_j(t)))/\eta_j$ does not depend on j and is called the **throughput** of this network. We denote the throughput $G(N-1,J)/G(N,J)$ of a Gordon-Newell network by

$$TH(R/\mu/J+N).$$

It is interesting to compare throughput for two structured networks with different routing and/or service properties. We shall present such results later in this chapter.

8.2 Stochastic monotonicity and related properties for classical networks

It is remarkable that many stochastic processes possess in a natural way some stochastic monotonicity properties. Among them, for example birth and death processes, attractive particle systems, and many classical queueing networks, which are in a sense similar to birth and death processes but more general because of migrations (movements to non-comparable states). It is not clear how the celebrated product form stationary distribution for networks is related to the property of stochastic monotonicity. There are product form networks which are not stochastically monotone, and there are stochastically monotone networks which are not in the class of product form networks. However it is not surprising that stochastically monotone networks which are at the same time product form networks possess many interesting properties. These both properties allow for many interesting comparison results and consequently also for many dependency results (for example dependency orderings). Stochastic monotonicity can have different forms depending on the ordering we select in the state space, and the shape of a network (for example for tandems we have so called partial sum monotonicity). Such properties will be the main topic of this section. For more general networks the area of monotonicity still remains to a large extend an open area.

8.2.1 Stochastic orders and monotonicity

From a general point of view, we shall consider probability measures on a partially ordered Polish space $\mathbb{E}$ endowed with a closed partial order $\prec$, and the Borel

$\sigma-$algebra $\mathcal{E}$ denoted by $\mathbb{E}$ along with random elements $X : \Omega \to \mathbb{E}$. We denote by $\mathcal{I}^*(\mathbb{E})$ ($\mathcal{I}^*_+(\mathbb{E})$) the set of all real valued increasing measurable bounded (non-negative) functions on $\mathbb{E}$ (f increasing means: for all $x, y, x \prec y$ implies $f(x) \le f(y)$), and $\mathcal{I}(\mathbb{E})$ the set of all increasing sets (i.e. sets for which indicator functions are increasing). The *decreasing* analogues are denoted by $\mathcal{D}^*(\mathbb{E})$, ($\mathcal{D}^*_+(\mathbb{E})$) and $\mathcal{D}(\mathbb{E})$, respectively. For $A \subseteq \mathbb{E}$ we denote $A^\uparrow := \{y \in \mathbb{E} : y \succ x \ for\ some\ x \in A\}$, and $A^\downarrow := \{y \in \mathbb{E} : y \prec x \ for\ some\ x \in A\}$. Further, we define $\mathcal{I}_p(\mathbb{E}) = \{\{x\}^\uparrow : x \in \mathbb{E}\}$ and $\mathcal{D}_p(\mathbb{E}) = \{\{x\}^\downarrow : x \in \mathbb{E}\}$, the classes of *one-point generated increasing*, resp. decreasing, sets.

For product spaces we shall use the following notation, $\mathbb{E}^{(n)} = \mathbb{E}_1 \times ... \times \mathbb{E}_n$, for $\mathbb{E}_i$ partially ordered Polish spaces ($i = 1, \dots, n$). If $\mathbb{E}_i = \mathbb{E}$ for all i then we write $\mathbb{E}^n$ instead of $\mathbb{E}^{(n)}$. Analogously we write $\mathbb{E}^{(\infty)}$ and $\mathbb{E}^\infty$ for infinite products. Product spaces will be considered with the product topology. Elements of $\mathbb{E}^{(n)}$ will be denoted by $x^{(n)} = (x_1, \dots, x_n)$, of $\mathbb{E}^{(\infty)}$ by $x^{(\infty)}$. For random elements we use capital letters in this notation. We denote the coordinatewise ordering on $\mathbb{E}^{(n)}$ by $\prec^{(n)}$.

The theory of dependence order via integral orders for finite dimensional vectors is well established, surveys can be found in Mueller and Stoyan [65], Joe [37], and Szekli [88]. In recent years this theory and its applications were extended to dependence order of stochastic processes, see for examples with state spaces $\mathbb{R}^n$ or subsets thereof, the work of Hu and Pan [34] and Li and Xu [55], and for a more general approach to Markov processes in discrete and continuous time with general partially ordered state space, Daduna and Szekli [24].

Definition 8.2.1 *We say that two random elements* $\mathbf{X}, \mathbf{Y}$ *of* $\mathbb{E}$ *are stochastically ordered (and write* $\mathbf{X} \prec_{st} \mathbf{Y}$ *or* $\mathbf{Y} \succ_{st} \mathbf{X}$*) if* $Ef(\mathbf{X}) \le Ef(\mathbf{Y})$ *for all* $f \in \mathcal{I}^*(\mathbb{E})$*, for which the expectations exist.*

In the theory of stochastic orders and especially in specific applications a well established procedure is to tailor suitable classes of functions, which via integrals over these functions extract the required properties of the models under consideration. The most well known example is the class of integrals over convex functions which describes the volatility of processes and therefore the risks connected with their evolution.

Similar ideas will guide our investigations of network processes $\mathbf{X} = (X_t : t \ge 0)$ and $\mathbf{Y} = (Y_t : t \ge 0)$. These are comparable in the concordance ordering, $\mathbf{X} \prec_{cc} \mathbf{Y}$, if for each pair $(X_{t_1}, \dots, X_{t_n})$ and $(Y_{t_1}, \dots, Y_{t_n})$ it holds

$$E\left[\prod_{i=1}^{n} f_i(X_{t_i})\right] \le E\left[\prod_{i=1}^{n} f_i(Y_{t_i})\right], \tag{8.6}$$

for all increasing functions f_i, and for all decreasing functions as well (i.e. for all comonotone functions). It is our task to identify subclasses $\mathcal{F}$ of functions such that (8.6) holds for all comonotone functions which are additionally in $\mathcal{F}$ and that additionally $\mathbf{X}$ and $\mathbf{Y}$ fulfill the corresponding stochastic monotonicity properties with respect to the integral order defined via intersecting the class of monotone functions with $\mathcal{F}$.

The set (8.6) of inequalities implies that $\mathbf{X}$ and $\mathbf{Y}$ have the same marginals and that standard covariances $cov(f(X_s),g(X_t)) \leq cov(f(Y_s),g(Y_t))$ are ordered for all comonotone f,g. If $\mathcal{F}$ is sufficiently rich, these properties will be maintained.

The introduced above dependence ordering can be generalized for more general spaces. For $\mathbb{E}$ which is a lattice (i.e. for any $x,y \in \mathbb{E}$ there exist a largest lower bound $x \wedge y \in \mathbb{E}$ and a smallest upper bound $x \vee y \in \mathbb{E}$ uniquely determined) we denote by $\mathcal{L}_{sm}(\mathbb{E})$ the set of all real valued bounded measurable supermodular functions on $\mathbb{E}$, i.e., functions which fulfill for all $x,y \in \mathbb{E}$

$$f(x \wedge y) + f(x \vee y) \geq f(x) + f(y).$$

Definition 8.2.2 *We say that two random elements* $\mathbf{X},\mathbf{Y}$ *of* $\mathbb{E}$ *are supermodular stochastically ordered (and write* $\mathbf{X} \prec_{sm} \mathbf{Y}$ *or* $\mathbf{Y} \succ_{sm} \mathbf{X}$*) if* $Ef(\mathbf{X}) \leq Ef(\mathbf{Y})$ *for all* $f \in \mathcal{L}_{sm}(\mathbb{E})$, *for which the expectations exist.*

A weaker than $\prec_{sm}$ can be defined on product spaces. A function $f : \mathbb{E}^{(2)} \to \mathbb{R}$ has *isotone differences* if for $x_1 \prec_1 x_1'$, $x_2 \prec_2 x_2'$ we have

$$f(x_1',x_2') - f(x_1,x_2') \geq f(x_1',x_2) - f(x_1,x_2). \tag{8.7}$$

A function $f : \mathbb{E}^{(n)} \to \mathbb{R}$ has *isotone differences* if (8.7) is satisfied for any pair i,j of coordinates, whereas the remaining variables are fixed. If $\mathbb{E}_i$, $i = 1,\dots,n$ are totally ordered then both definitions are equivalent. The class of functions with isotone differences, defined by (8.7), we denote by $\mathcal{L}_{\text{idif}}(\mathbb{E}^{(n)})$. Note that the definition of a function with isotone differences does not require that $\mathbb{E}_i$ are lattices. If, additionally, f is taken to be increasing we shall write $f \in \mathcal{L}_{\text{i-idif}}(E^{(n)})$. The following lemma is due to Heyman and Sobel [44].

Lemma 8.2.3 (i) *Let* $\mathbb{E}_1,\mathbb{E}_2,\dots,\mathbb{E}_n$ *be lattices. If* f *is supermodular on* $(\mathbb{E}^{(n)},\prec^{(n)})$ *then it has also isotone differences.*
(ii)*Let* $\mathbb{E}_1,\dots,\mathbb{E}_n$ *be totally ordered. If* f *has isotone differences on* $(\mathbb{E}^{(n)},\prec^{(n)})$ *then it is also supermodular.*

The above lemma implies that for totally ordered spaces both notions are equivalent. This is not the case when $\mathbb{E}_i$, $i = 1\dots,n$, are partially (but not linearly) ordered.

Definition 8.2.4 *Let* $\mathbf{X} = (X_1,\dots,X_n)$, $\mathbf{Y} = (Y_1,\dots,Y_n)$ *be random vectors with values in* $\mathbb{E}^{(n)}$.

(i) $\mathbf{X}$ *is smaller than* $\mathbf{Y}$ *in the isotone differences ordering (*$\mathbf{X} \prec_{idif} \mathbf{Y}$*) if*

$$E\left[f(X_1\dots,X_n))\right] \leq E\left[f(Y_1,\dots,Y_n)\right]$$

for all $f \in \mathcal{L}_{\text{idif}}(E^{(n)})$.

Let us summarize some definitions which we will need later.

8.2.1.1 Discrete time

Let $\mathbf{X} = (X_t : t \in \mathbb{Z})$ and $\mathbf{Y} = (Y_t : t \in \mathbb{Z}), X_t, Y_t : \Omega \to \mathbb{E}$, be discrete time, stationary, homogeneous Markov processes. Assume that π is an invariant (stationary) one–dimensional marginal distribution the same for both $\mathbf{X}$ and $\mathbf{Y}$, and denote the 1–step transition kernels for $\mathbf{X}$ and $\mathbf{Y}$, by $K^X : \mathbb{E} \times \mathcal{E} \to [0,1]$, and $K^Y : \mathbb{E} \times \mathcal{E} \to [0,1]$, respectively. Denote the respective transition kernels for the time reversed processes $\overleftarrow{\mathbf{X}}$, $\overleftarrow{\mathbf{Y}}$ by $\overleftarrow{K^X}$, $\overleftarrow{K^Y}$. We say that a stochastic kernel $K : \mathbb{E} \times \mathcal{E} \to [0,1]$ is **stochastically monotone** if $\int f(x)K(s,dx)$ is increasing in s, for each $f \in \mathcal{I}^*(\mathbb{E})$. It is known (see e.g. Müller and Stoyan [65], section 5.2) that a stochastic kernel K is stochastically monotone iff $K(x,\cdot) \prec_{st} K(y,\cdot)$ for all $x \prec y$. Another equivalent condition for this property is that $\mu K \prec_{st} \nu K$, for all $\mu \prec_{st} \nu$, where μK denotes the measure defined by $\mu K(A) = \int K(s,A)\mu(ds), A \in \mathcal{E}$. It is worth mentioning that for $\mathbb{E} = \mathbb{N}$, using traditional notation $P^X = [p^X(i,j)]_{i,j \in \mathbb{N}}$ for the transition matrix of $\mathbf{X}$ (that is $p^X(i,j) := K^X(i,\{j\})$), stochastic monotonicity can be expressed in a very simple form, namely (see Keilson and Kester [39]), we say that P^X is stochastically monotone iff

$$T^{-1}P^X T(i,j) \geq 0, \; i,j \in \mathbb{N}, \tag{8.8}$$

where T is the lower triangular matrix with zeros above the main diagonal and ones elsewhere.

8.2.1.2 Continuous time

Let $\mathbf{X} = (X_t : t \in \mathbb{R})$ and $\mathbf{Y} = (Y_t : t \in \mathbb{R}), X_t, Y_t : \Omega \to \mathbb{E}$, be stationary homogeneous Markov processes. Denote the corresponding families of transition kernels of $\mathbf{X}$, and $\mathbf{Y}$, by $\mathbb{K}^X = (K_t^X : \mathbb{E} \times \mathcal{E} \to [0,1] : t \geq 0)$, and $\mathbb{K}^Y = (K_t^Y : \mathbb{E} \times \mathcal{E} \to [0,1] : t \geq 0)$, respectively, and the respective transition kernels for the stationary time reversed processes $\overleftarrow{\mathbf{X}}$, $\overleftarrow{\mathbf{Y}}$ by $\overleftarrow{\mathbb{K}^X} = (\overleftarrow{K_t^X} : \mathbb{E} \times \mathcal{E} \to [0,1] : t \geq 0)$, and $\overleftarrow{\mathbb{K}^Y} = (\overleftarrow{K_t^Y} : \mathbb{E} \times \mathcal{E} \to [0,1] : t \geq 0)$, respectively. Assume that π is an invariant distribution common for both $\mathbb{K}^X$ and $\mathbb{K}^Y$, that is $\int K_t^X(x,dy)\pi(dx) = \int K_t^Y(x,dy)\pi(dx) = \pi(dy)$, for all $t > 0$. We say that $\mathbb{K}^X$ ($\mathbb{K}^Y$) is **stochastically monotone** if for each $t > 0$, K_t^X (K_t^Y) is stochastically monotone as defined previously. If $\mathbb{E}$ is countable and $Q^{\mathbf{X}} = [q^{\mathbf{X}}(x,y)]$ and $Q^{\mathbf{Y}} = [q^{\mathbf{Y}}(x,y)]$ denote intensity matrices (infinitesimal generators) of the corresponding chains $\mathbf{X}$ and $\mathbf{Y}$ then the following condition due to Massey [60] is useful: if $Q^{\mathbf{X}}$ is bounded, conservative then $\mathbb{K}^X$ is stochastically monotone iff

$$\sum_{y \in F} q^{\mathbf{X}}(x_1,y) \leq \sum_{y \in F} q^{\mathbf{X}}(x_2,y),$$

for all $F \in \mathcal{I}(\mathbb{E})$, and $x_1 \prec x_2$ such that $x_1 \in F$ or $x_2 \notin F$. An analogous condition for arbitrary time continuous Markov jump processes (also for unbounded generators) is given by Mu-Fa Chen [15], Theorem 5.47. It is worth mentioning that if $\mathbb{E} = \mathbb{N}$ then similarly to (8.8), we say that $Q^{\mathbf{X}} = [q^{\mathbf{X}}(i,j)]_{i,j \in \mathbb{N}}$ is stochastically monotone iff $T^{-1}Q^{\mathbf{X}}T(i,j) \geq 0$ for all $i \neq j$.

8.2.2 Stochastic monotonicity and networks

The fundamental property of stochastic monotonicity of classical networks is presently very well known. Massey [60], Proposition 8.1, proved this property using analytical methods for Jackson networks with constant service rates, Daduna and Szekli [18], Corollary 4.1, utilized a coupling argument combined with a point processes description, admitting variable service rates and closed networks. Lindvall [51], p. 7, used a coupling proof for Jackson networks.

Property 8.2.5 *Consider* $\tilde{\mathbf{X}} := (\tilde{X}(t), t \geq 0)$ *the joint queue length process in a Jackson network* $(\lambda, \tilde{R}/\mu/J)$ *as a Markov process with the partially ordered state space* $(\mathbb{E}, \prec) := (\mathbb{N}^J, \leq^J)$*, and* $\mathbf{X} = (X(t) : t \geq 0)$ *the process recording the joint queue lengths in the Gordon-Newell network* $(R/\mu/J+N)$ *as a Markov process with state space* $\mathbb{E}_N$*, also ordered with* $\leq^J$ *(the standard coordinate-wise ordering). If* μ *is increasing as a function of the number of customers then for both processes the corresponding families of transition kernels are stochastically monotone with respect to* $\leq^J$.

Remark 8.1. For a formulation of the above result in terms of marked point processes see Last and Brandt [49], Theorem 9.3.18. For a version of the stochastic monotonicity property for Jackson networks with infinite denumerable number of nodes see Kelbert et al. [52]. For a refined stochastic monotonicity property, for partition separated orderings, see Proposition 8.1 in Massey [60]. For generalizations to Jackson type networks with batch movements see Economou [26] and [27].

Apart from the traditional, coordinatewise ordering on the state space it is possible and reasonable to consider other orderings and monotonicities which for example turned out to be useful to describe special properties of tandems.

For two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\mathbf{x} = (x_1, \ldots, x_n)$, $\mathbf{y} = (y_1, \ldots, y_n)$, we define **partial sum** order by

$$\mathbf{x} \leq_* \mathbf{y} \text{ if } \sum_{i=1}^{j} x_i \leq \sum_{i=1}^{j} y_j, \qquad j = 1, \ldots, n.$$

The next property was first stated by Whitt [97], and restated using other methods in Massey [60], Theorem 8.3, and Daduna and Szekli [18], Proposition 4.4.

Property 8.2.6 *Consider* $\tilde{\mathbf{X}} := (\tilde{X}(t), t \geq 0)$ *the joint queue length process in Jackson network* $(\lambda, \tilde{R}/\mu/J)$ *as a Markov process with the partially ordered state space* $(\mathbb{E}, \prec) := (\mathbb{N}^J, \leq_*)$. *Assume that* μ *is increasing as a function of the number of customers. Then the corresponding family of transition kernels of* $\tilde{\mathbf{X}}$ *is stochastically monotone with respect to* $\leq_*$ *if and only if* $i, j \in \mathrm{J}$ *and* $\tilde{r}(i,j) > 0$ *implies that* $j = i+1$ *or* $j = i-1$*, and* $\tilde{r}(i,0) > 0$ *iff i=J.*

An interesting monotonicity property for increments of cumulative number of customers in Jackson networks starting empty was proved by Lindvall [51] using coupling methods.

Property 8.2.7 *Consider* $\tilde{\mathbf{X}}$ *the joint queue length process in Jackson network* $(\lambda,\tilde{R}/\mu/J)$ *such that at time 0 the system is empty. Assume that* μ *is increasing as a function of the number of customers. Then for each* $\varepsilon > 0$, $\sum_{j=1}^{J}\tilde{X}_j(t+\varepsilon) - \sum_{j=1}^{J}\tilde{X}_j(t)$ *is stochastically* ($\leq_{st}$) *decreasing as a function of* t.

8.2.3 Bounds in transient state

The analytical approach of Massey [58], [59] resulted in a transient bound for Jackson networks which was generalized then by Tsoucas and Walrand [93]. The joint distribution of the number of customers on an upper orthant can be bounded from above by the product of the corresponding state distributions of single systems at any time provided they start from the same state. This is a useful upper bound on the probability of overload in transient Jackson networks.

Property 8.2.8 *Consider* $\tilde{\mathbf{X}}$ *the joint queue length process in Jackson network* $(\lambda,\tilde{R}/\mu/J)$ *such that* $\mu = (\mu_1,\ldots,\mu_J)$ *is constant as a function of the number of customers. Independently, for each* $j \in$ J, *denote by* $X_j^*(t)$ *the number of customers in the* $M/M/1$*-FCFS classical system with the arrival rate*

$$\lambda_j^* = \tilde{r}_{0j}\lambda + \sum_{i=1}^{J}\mu_i\tilde{r}_{ij},$$

and the service rate μ_j. *If for the initial conditions* $\tilde{X}(0) = (X_1^*(0),\ldots,X_J^*(0)$ *then*

$$P(\tilde{X}(t) \geq a) \leq P(X_1^*(t) \geq a_1)\cdots P(X_j^*(t) \geq a_j),$$

for each $t > 0$ *and* $a = (a_1,\ldots,a_j) \in \mathbb{R}^J$.

8.2.4 Bounds in stationary state

Bounds for time stationary evolution of networks have a different nature than transient bounds. The next property can be found in Daduna and Szekli [18], Corollary 5.1.

Property 8.2.9 *Consider* $\tilde{\mathbf{X}}$ *the joint queue length process in Jackson network* $(\lambda,\tilde{R}/\mu/J)$ *such that* $\mu = (\mu_1(\cdot),\ldots,\mu_J(\cdot))$ *is increasing as a function of the number of customers. Then in stationary conditions*

$$E(f[\tilde{X}(t_1),\ldots,\tilde{X}(t_i)]g[\tilde{X}(t_{i+1}),\ldots,\tilde{X}(t_k)])$$

$$\geq E(f[\tilde{X}(t_1),\ldots,\tilde{X}(t_i)])E(g[\tilde{X}(t_{i+1}),\ldots,\tilde{X}(t_k)])$$

for all non-decreasing real f,g, *and* $0 \le t_1 < \ldots < t_k, i < k,\ \ i,k \in \mathbb{N}$.

This inequality can be written as

$$\mathrm{Cov}(f(\tilde{X}(t_i), i = 1,\ldots,k), g(\tilde{X}(t_i), i = k+1,\ldots,n)) \ge 0, \tag{8.9}$$

for all $f \in \mathcal{I}^*(\mathbb{R}^k), g \in \mathcal{I}^*(\mathbb{R}^{n-k}), k = 1,\ldots,n-1, t_1 < \ldots < t_n$. Note that the property (8.9) is a rather strong positive dependence property in the time evolution of $\tilde{\mathbf{X}}$. We shall recall now some definitions from the theory of positive dependence. A natural way to define positive dependence for a random vector (or alternatively for a distribution on a state space) $\mathbf{X} = (X_1,\ldots,X_n)$ is to use a dependency ordering in order to compare it with its iid version, i.e. with $\mathbf{X}^\perp = (X_1^\perp,\ldots,X_n^\perp)$, where $X_i =^d X_i^\perp$, and $(X_1^\perp,\ldots,X_n^\perp)$ being independent. For example, if $\mathbb{E} = \mathbb{R}$, $\mathbf{X}^\perp \le_{cc} \mathbf{X}$ is equivalent to the fact that $\mathbf{X}$ is positively orthant dependent (POD) (for definitions of this and other related concepts see e.g. Szekli [88]). POD is weaker than association of $\mathbf{X}$ defined by the condition that $\mathrm{Cov}(f(\mathbf{X}), g(\mathbf{X})) \ge 0$ for all $f,g \in \mathcal{I}^*(\mathbb{R}^n)$. However, it is not possible to characterize association in terms of some ordering, that is by stating that $\mathbf{X}$ is greater than $\mathbf{X}^\perp$ for some ordering. But Christofides and Veggelatou [17] show that association implies that $\mathbf{X}^\perp \le_{sm} \mathbf{X}$ (positive supermodular dependence - PSMD). In fact they show that the weak association (defined by $\mathrm{Cov}(f(X_i, i \in A), g(X_i, i \in A^c)) \ge 0$ for all real, increasing f,g of appropriate dimension, and all $A \subset \{1,\ldots,n\}$) implies PSMD. Rüschendorf [74] defined a weaker than weak association positive dependence by $\mathrm{Cov}(I_{(X_i > t)}, g(X_{i+1},\ldots,X_n)) \ge 0$ for all increasing g, all $t \in \mathbb{R}$, and all $i = 1,\ldots,n-1$, which he called weak association in sequence (WAS). He showed that WAS implies PSMD. Hu et al. [35] gave counterexamples showing that the mentioned positive dependence concepts are really different.

Note that property (8.9) implies that $(f_1(\tilde{X}(t_1)),\ldots,f_n(\tilde{X}(t_n)))$ is weakly associated in sequence for all $f_i \in \mathcal{I}_+^*(\mathbb{R}^J)$, and therefore is also PSMD, which implies possibility to compare maxima, minima and other supermodular functionals of the time evolution of $\tilde{\mathbf{X}}$, $(f_1(\tilde{X}(t_1)),\ldots,f_n(\tilde{X}(t_n)))$ with the corresponding independent versions (separated single queue systems). This is in accordance with intuitions since the joint time evolution of a network should generate more correlations than independent single queue systems.

It is worth mentioning that in order to obtain a joint space and time positive dependence for a Markov process $\mathbf{X}$ one requires additional assumptions. For example it is known (see e.g. Liggett [50], Szekli [88], Theorem A, section 3.7.) that if $I\!\!K^X$ is stochastically monotone, π associated on $\mathbb{E}$, and (so called up-down property) $Q^{\mathbf{X}}(fg) \ge fQ^{\mathbf{X}}g + gQ^{\mathbf{X}}f$, for all f,g increasing then $\mathbf{X}$ is space-time associated (i.e. $\mathrm{Cov}(\phi(X_{t_i}, i = 1,\ldots,n), \psi(X_{t_i}, i = 1,\ldots,n)) \ge 0$, for all ϕ, ψ increasing). Unfortunately networks in general do not fulfill this up-down requirement therefore the last property needed another argument strongly based on stochastic monotonicity.

The next property is a corollary from the previous one but it is interesting to know that it is possible to extend this property to networks of infinite channel queues with arbitrary service time distribution, see Kanter [38], Daduna and Szekli [18], Corollary 5.2. In contrast to the transient case these bounds are lower bounds and are formulated with respect to the time evolution in stationary conditions.

Property 8.2.10 *Consider* $\tilde{\mathbf{X}}$ *the joint queue length process in Jackson network* $(\lambda, \tilde{R}/\mu/J)$ *such that* $\mu = (\mu_1(\cdot), \ldots, \mu_J(\cdot))$ *is increasing as a function of the number of customers. Independently, for each* $j \in \mathrm{J}$*, denote by* $X_j^*(t)$ *the number of customers in the* $M/M(n)/1$*-FIFO classical system with the arrival rate* $\lambda_j^* = \tilde{\eta}_j$ *and the service rate* $\mu_j(\cdot)$*. Then for both processes in stationary conditions*

$$P(\tilde{X}(t_1) \geq (\leq) a^1, \ldots, \tilde{X}(t_k) \geq (\leq) a^k) \geq \prod_{1 \leq i \leq k, 1 \leq j \leq J} P(X_j^*(t_i) \geq (\leq) a_j^i),$$

for each $t_1 < \cdots < t_k$*, and* $a^k = (a_1^k, \ldots, a_j^k) \in \mathbb{R}^J$*,* $k \in \mathbb{N}$.

For open networks in stationary state positive correlations are prevailing. For closed networks however it is natural to expect negative correlations for the state in closed networks, but negative association is perhaps a bit surprising at the first glance. The next property can be found in Daduna and Szekli [18], Proposition 5.3.

Property 8.2.11 *Consider* $\mathbf{X} = (X(t) : t \geq 0)$ *the process recording the joint queue lengths in the Gordon-Newell network* $(R/\mu/J+N)$ *as a Markov process with state space* $\mathbb{E}_N$ *ordered with* $\leq^J$*. If* μ *is increasing as a function of the number of customers then for every* $t > 0$*,* $X(t)$ *is negatively associated with respect to* $\leq^J$*, i.e.*

$$E(f(X_i(t), i \in \mathrm{I}) g(X_j(t), j \in \mathrm{I}^c)) \leq E(f(X_i(t), i \in \mathrm{I})) E(g(X_j(t), j \in \mathrm{I}^c)),$$

for all increasing f, g*, and all* $\mathrm{I} \subset \mathrm{J}$.

For analogous result for discrete time queueing networks see Pestien and Ramakrishnan [69]. Negative association can be used to obtain upper bounds on the joint distribution of the state vector.

8.2.5 Sojourn times in networks

8.2.5.1 Dependence properties for sojourn times

A path of length M in the network $(\lambda, \tilde{R}/\mu/J)$ is a finite sequence of nodes $\mathcal{P} = (j_1, j_2, \ldots, j_M)$, not necessarily distinct, which a customer can visit consecutively, i.e., $\tilde{r}_{j_k, j_{k+1})} > 0, k = 1, \ldots, M-1$. For a customer traversing path $\mathcal{P}$ we denote by $(\xi_1, \xi_2, \ldots, \xi_M)$ the vector of his successive sojourn times at the nodes of the path. Strong interest is focused on determining the joint distribution of the vector $\xi = (\xi_1, \ldots, \xi_N)$ in equilibrium. In general this is an unsolved problem, explicit expressions are rare.

The first results were obtained by Reich [71], [72], and Burke [9], [10]. For closed cycles the parallel results were developed by Chow [16], Schassberger and Daduna

[85], and Boxma, Kelly, and Konheim [8]. Clearly in this case independence was not found due to the negative correlation of the queue lengths in the network, but a product form structure emerged there as well. The research which followed the mentioned early results was also concentrated on proving that similar results hold for *overtake–free* paths as well. Extensions to single server overtake–free paths for networks with general topology were obtained for the open network case by Walrand and Varaiya [95] and Melamed [63], and for closed networks by Kelly and Pollett [41]. The result for overtake–free paths with multiserver stations at the beginning and the end of the path was proved by Schassberger and Daduna [86]. (For a review see Boxma and Daduna [7].)

The most prominent example where overtaking appears is the Simon–Foley [87] network of single server queues, see Fig. 8.1. As we have already mentioned before, the question whether on the three–station path of the Simon–Foley network the complete sojourn time vector (ξ_1, ξ_2, ξ_3) is associated remains unanswered. The methods provided by the proof of Foley and Kiessler [28] seemingly do not apply to that problem. However it is possible to prove a little bit weaker dependence results. Probability measure used in this statement is the Palm probability with respect to the point process of arrivals to the first station.

Property 8.2.12 *Consider Jackson network* $(\lambda, \tilde{R}/\mu/J)$ *with constant* μ*, and a path* $\mathcal{P}$ *consisting of three nodes which we assume to be numbered* $\mathcal{P} = (1,2,3)$*. In equilibrium, the successive sojourn times* (ξ_1, ξ_2, ξ_3) *of a customer on a three node path of distinct nodes are positive upper orthant dependent, i.e.*

$$P(\xi_1 \geq a_1, \xi_2 \geq a_2, \xi_3 \geq a_3) \geq P(\xi_1 \geq a_1)P(\xi_2 \geq a_2)P(\xi_3 \geq a_3)$$

.

More generally the above result holds true in open product form networks with multi-server nodes having general service disciplines and exponentially distributed service times or having symmetric service disciplines with generally distributed service times. Moreover this is true also for networks with customers of different types entering the network and possibly changing their types during their passage through the network. Here one may allow additionally that the service time distributions at symmetric nodes are type dependent, see Daduna and Szekli [19]. For generalizations to four step walk in Jackson networks see Daduna and Szekli [20].

8.2.5.2 Sojourn times in closed networks

Intuitively, sojourn times in closed networks should be negatively correlated, but again negative association is a bit surprising as a property explaining this intuition. The next property for closed cycles of queues is taken from Daduna and Szekli [21]. The expectations in this statement are taken with respect to the Palm measure defined with respect to the point process of transitions between two fixed consecutive stations.

Property 8.2.13 *Consider Gordon-Newell network $(R/\mu/J+N)$ with constant μ, and cyclic structure of transitions, i.e. $r_{i(i+1)} = 1$ for $i \le J-1$ and $r_{J1} = 1$. In equilibrium, for the successive sojourn times $(\xi_1,\dots,\xi_J)$ of a customer at stations $1,\dots,J$,*

$$E(f(\xi_i, i \in \mathrm{I})g(\xi_j, j \in \mathrm{I}^c)) \le E(f(\xi_i, i \in \mathrm{I}))E(g(\xi_j, j \in \mathrm{I}^c)),$$

for all increasing f,g, and all $\mathrm{I} \subset \mathrm{J}$, i.e. ξ is negatively associated.

In a closed tandem system with fixed population size the conditional cycle time distribution of a customer increases in the strong stochastic ordering when the initial disposition of the other customers increases in the partial sum ordering. As a consequence of this property one obtains

Property 8.2.14 *Consider Gordon-Newell network $(R/\mu/J+N)$ with constant μ, and cyclic structure of transitions, i.e. $r_{i(i+1)} = 1$ for $i \le J-1$ and $r_{J1} = 1$. In equilibrium, the cycling time $\xi_1 + \dots + \xi_J$ of a customer going through stations $1,\dots,J$ is stochastically increasing in N, the number of customers cycling.*

For negative association (NA) of sojourn times in the consecutive cycles made by a customer, see Daduna and Szekli [21].

8.3 Properties of throughput in classical networks

8.3.1 Uniform conditional variability ordering, a relation between closed and open networks

The next property is taken from Whitt [98]. Before formulating it we need some definitions.

Definition 8.3.1 *Suppose that μ, ν are probability measures which are not related by the stochastic ordering $\le_{st}$, and are absolutely continuous with respect to Lebesque (counting) measure on $\mathbb{R}$ ($\mathbb{N}$) with densities (mass functions) f, g respectively, with $supp(\mu) \subset supp(\nu)$. We say that*

1. *μ is uniformly conditionally less variable than ν, and write $\mu \prec_{uv} \nu$ if $f(t)/g(t)$ is unimodal on $t \in supp(\nu)$, with the mode being the supremum.*
2. *μ is log-concave relative to ν, and write $\mu \prec_{lcv} \nu$ if $supp(\mu) \subset supp(\nu)$ are intervals (connected sets of integers) and $f(t)/g(t)$ is log-concave on $t \in supp(\mu)$.*
3. *$\mu \prec_{mlr} \nu$ if $f(t)/g(t)$ is nonincreasing on $t \in supp(\mu)$.*

We use also $\prec_{lcv}$ and $\prec_{uv}$ to relate random variables using the above definition for their distributions.

If the number of sign changes $S(f-g) = 2$, and $\mu \prec_{lcv} \nu$ then $\mu \prec_{uv} \nu$. Moreover if $\mu(A), \nu(B) > 0, A \subset B, S(f-g) = 2$, and $\mu \prec_{lcv} \nu$ then

(i) if $E(\mu_A) \le E(\nu_B)$ then $\mu_A \le_{icx} \nu_B$
(ii)if $E(\mu_A) \ge E(\nu_B)$ then $\mu_A \le_{dcx} \nu_B$
(iii)$E(\mu_A) = E(\nu_B)$ then $\mu_A \le_{cx} \nu_B$,

where $E(\mu)$ denotes the expected value of μ, and μ_A denotes the conditional distribution of μ conditioned on A.

It is known (see Whitt [98]) that for each Gordon-Newell network $(R/\mu/J+N)$ there exist a Jackson network $(\lambda, \tilde{R}/\mu/J)$, such that the stationary distribution of the network content in Gordon-Newell model is equal to the conditional stationary distribution in this Jackson model, conditioned on the fixed number of customers, that is $\pi^{(N,J)}(n) = \tilde{\pi}^J(n \mid \{n : \sum_{i=1}^J n_i = N\})$. For each such pair of stationary network processes $\mathbf{X}, \tilde{\mathbf{X}}$ it is possible to compare variability of the corresponding one dimensional marginal distributions if for each i, $\mu_i(n)$ are nondecreasing functions of n.

Property 8.3.2 *In stationary conditions it holds that for all t*

$$X_i(t) \prec_{lcv} \tilde{X}_i(t),\ i = 1,\dots,J.$$

From the above relation it follows that if $E(\sum_{i=1}^J \tilde{X}_i(t)) \le N$ then for respective utilizations at each node i

$$E(\tilde{X}_i(t) \wedge s_i) \le E(X_i(t) \wedge s_i),$$

provided $\mu_i(n) = (n \wedge s_i)\mu$ for some $s_i \in \mathbb{N}$, and $\mu > 0$ or equivalently for ***throughputs***

$$E(\mu_i(\tilde{X}_i(t))) \le E(\mu_i(X_i(t))).$$

8.3.2 Effect of enlarging service rates in closed networks

Chen and Yao [14] showed that if in a closed network, locally in some set of nodes the service rates will be increased then the number of customers in these nodes will decrease, but the number of customers elsewhere will increase (in $\prec_{mlr}$ sense). Moreover the overall throughput for the network will be larger.

Property 8.3.3 *Suppose that we consider two Gordon-Newell networks $(R/\mu/J+N)$ and $(R/\mu'/J+N)$, and the corresponding stationary queue length processes $\mathbf{X}$, $\mathbf{X}'$, such that for a subset $A \subset \{1,\dots,J\}$, $\mu_j \le \mu'_j$ (pointwise) for $j \in A$, and $\mu_j = \mu'_j$, for $j \in A^c$. Then*

$$X'_j(t) \prec_{mlr} X_j(t)$$

for $j \in A$ and

$$X_j(t) \prec_{mlr} X'_j(t)$$

for $j \in A^c$. Moreover if $\mu_j(n)$ are nondecreasing functions of n then

$$TH(R/\mu/J+N) \leq TH(R/\mu'/J+N).$$

From Shanthikumar and Yao [81] we have

Remark 8.2. If we change the condition that $\mu_j \leq \mu'_j$ (pointwise) for $j \in A$, by a stronger one: $\mu_j(m) \leq \mu'_j(n)$ for $j \in A$, and all $m \leq n$, $m,n \in \mathbb{N}$ then $TH(R/\mu/J+N) \leq TH(R/\mu'/J+N)$ holds without assuming monotonicity of service rates. We have for example $TH(R/\mu_{min}/J+N) \leq TH(R/\mu/J+N) \leq TH(R/\mu_{max}/J+N)$, whenever $\mu_{min} = (\min_{n\geq 1}\mu_1(n),\ldots,\min_{n\geq 1}\mu_J(n))$
and $\mu_{max} = (\max_{n\geq 1}\mu_1(n),\ldots,\max_{n\geq 1}\mu_J(n))$ are finite, positive.

8.3.3 Majorization, arrangement and proportional service rates

For two vectors $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$ we define the relation $\mathbf{x} \prec_{\mathbf{m}} \mathbf{y}$ by

$$\sum_{i=1}^{k} x_{[i]} \leq \sum_{i=1}^{k} y_{[i]},\ k<n,\ \sum_{i=1}^{n} x_{[i]} = \sum_{i=1}^{n} y_{[i]}\ ,$$

where $x_{[1]} \geq \ldots \geq x_{[n]}$ denotes non-increasing rearrangement of $\mathbf{x}$. This relation is the **majorization**.

For two vectors $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$ such that $\mathbf{x}$ is a permutation of $\mathbf{y}$ we define the relation $\mathbf{x} \prec_{\mathbf{a}} \mathbf{y}$ by requiring that $\mathbf{y}$ can be obtained from $\mathbf{x}$ by a sequence of transpositions such that after transposition the two transposed elements are in decreasing order.

For the next properties in this subsection see Shanthikumar [78], and Chen and Yao [14]. The first one exploits interplay between some special regularities of the service rates (fulfilled for example for linear service rates) and a perturbation of the routing in such a way that after perturbation more probable are visits to the stations with lower numbers, which leads to a larger throughput. The second one again assumes some special properties for the service rates (proportional to increasing concave function), and non-increasing routing vector (more probable visits to the stations with lower numbers), then a perturbation leading to more decreasingly arranged service rates (more service for the stations with lower numbers) implies larger throughput.

Property 8.3.4 *Consider two Gordon-Newell networks* $(R/\mu/J+N)$, *and* $(R'/\mu/J+N)$ *such that all* $\mu_j(n)$ *are nondecreasing and concave in n, and* $\mu_j(n) - \mu_{j+1}(n)$ *is nondecreasing in n, for* $j \leq J-1$. *If for the corresponding routing probabilities* $\eta \prec_{\mathbf{a}} \eta'$ *then*

$$TH(R/\mu/J+N) \leq TH(R'/\mu/J+N).$$

Property 8.3.5 *Consider two Gordon-Newell networks* $(R/\mu/J+N)$, *and* $(R/\mu'/J+N)$ *such that for all j,* $\mu_j(n)$ *and* $\mu'_j(n)$ *are proportional* $\mu_j(n) = \mu_j c(n)$,

$\mu'_j(n) = \mu'_j c(n)$ to some $c(n)$ which is nondecreasing and concave in n, and η is non-increasing. If for the corresponding service intensities $\mu \prec_{\mathbf{a}} \mu'$ then

$$TH(R/\mu/J+N) \leq TH(R/\mu'/J+N).$$

Similar assumptions as above in the class of Jackson networks lead to the Schur-convex ordering of the state vectors, which here, intuitively speaking, describes a better performance of the network after the assumed perturbation (adjusting service capacities to the routing structure gives a better performance).

Property 8.3.6 *Consider two Jackson networks $(\lambda, \tilde{R}/\mu/J)$, and $(\lambda, \tilde{R}/\mu'/J)$ such that for all j, $\mu_j(n)$ and $\mu'_j(n)$ are proportional $\mu_j(n) = \mu_j c(n)$, $\mu'_j(n) = \mu'_j c(n)$ to some $c(n)$ which is nondecreasing and concave in n, and $\tilde{\eta}$ is non-increasing. If for the corresponding service intensities $\mu \prec_{\mathbf{a}} \mu'$ then*

$$E(\psi(\tilde{X}(t)) \geq E(\psi(\tilde{X}'(t))$$

for all nondecreasing and Schur-convex functions ψ.

The next property shows that if the vector of ratios: the probability of being in a station divided by its service intensity, has the property of being more equally distributed over the set of stations (in the sense of majorization) then it will lead to a larger throughput provided the service function is increasing and concave, and to smaller one if this function is increasing and convex.

Property 8.3.7 *Consider two Gordon-Newell networks $(R/\mu/J+N)$, and $(R'/\mu'/J+N)$ such that for all j, $\mu_j(n)$ and $\mu'_j(n)$ are proportional $\mu_j(n) = \mu_j c(n)$, $\mu'_j(n) = \mu'_j c(n)$ to some $c(n)$ which is nondecreasing and concave (convex) in n. If*

$$(\eta_1/\mu_1, \ldots, \eta_J/\mu_J) \prec_{\mathbf{m}} (\eta'_1/\mu'_1, \ldots, \eta'_J/\mu'_J)$$

then

$$TH(R/\mu/J+N) \geq (\leq) TH(R'/\mu'/J+N).$$

An analog of the above property can be formulated for Jackson networks.

Property 8.3.8 *Consider two Jackson networks $(\lambda, R/\mu/J)$, and $(\lambda', R'/\mu'/J)$ such that for all j, $\mu_j(n)$ and $\mu'_j(n)$ are proportional $\mu_j(n) = \mu_j c(n)$, $\mu'_j(n) = \mu'_j c(n)$ to some $c(n)$ which is nondecreasing and concave in n. If*

$$(\tilde{\eta}_1/\mu_1, \ldots, \tilde{\eta}_J/\mu_J) \prec_{\mathbf{m}} (\tilde{\eta}'_1/\mu'_1, \ldots, \tilde{\eta}'_J/\mu'_J)$$

then, in stationary conditions,

$$E(\psi(\tilde{X}(t)) \leq E(\psi(\tilde{X}'(t)),$$

for all nondecreasing and Schur-convex functions ψ.

A special case where for two networks the service rates are equal shows that the uniformly distributed routing gives the best throughput, if the service function is increasing and concave.

Property 8.3.9 *Consider two Gordon-Newell networks $(R/\mu/J+N)$, and $(R'/\mu/J+N)$ such that for all j, $\mu_j(n)$ are equal and nondecreasing and concave in n. If $\eta \prec_{\mathbf{m}} \eta'$ then*

$$TH(R/\mu/J+N) \geq TH(R'/\mu/J+N).$$

8.3.4 Throughput and number of jobs

Van der Wal [92] [1] obtained the following intuitively clear property

Property 8.3.10 *Suppose that for a Gordon-Newell network $(R/\mu/J+N)$ the service rates $\mu_i(n)$ are positive and nondecreasing functions of n, then in the stationary conditions $E(\mu_1(X_1(t)))$ is nondecreasing in N.*

From Chen and Yao [14], Shanthikumar and Yao [82], we have a more involved property.

Property 8.3.11 *Suppose that for a Gordon-Newell network $(R/\mu/J+N)$ the service rates $\mu_i(n)$ are positive and nondecreasing concave (convex, starshaped, antistarshaped, subadditive, superadditive) functions of n, then, in stationary conditions, $TH(R/\mu/J+N)$ has the same property treated as a function of N.*

The above property has an application to so called open - finite networks and blocking probabilities. Moreover Shanthikumar and Yao [83] studied monotonicity of throughput in cyclic/finite buffer networks with respect to the convex ordering of the service times, and of the buffer capacities.

8.4 Routing and correlations

General considerations on comparisons of Markov processes with respect to their internal dependence structure reveal that sometimes there is a complicated interplay of monotonicity properties with some generalized correlation structure of observed processes. Such monotonicity requirement is not unexpected if we recall that the theory of association in time for Markovian processes is mainly developed for monotone Markov processes, for a review see Chapter II of Liggett [50]. Association is a powerful tool in obtaining probability bounds e.g. in the realm of interacting processes of attractive particle systems. (A system is called attractive if it exhibits (strong) stochastic monotonicity.)

In the context of stochastic networks it turns out that similar connections between monotonicity and correlation are fundamental, but - due to a more complex structure of the processes we usually cannot hope to utilize the strong stochastic order, as required for association, or in the development by Hu and Pan [34], and Daduna and Szekli [24].

In this section we shall consider pairs of network processes related by some structural similarities. One can usually think of one network being obtained from the other by some structural perturbation. The perturbations we are mainly interested in are due to perturbing the routing of individual customers. We will always give a precise meaning of what the perturbations are and of the resulting structural properties. Proofs of all results presented in this section can be found in Daduna and Szekli [25].

We shall exhibit that the conditions that determine comparability of dependence, i.e., second order properties of processes having the same first order behavior (i.e. the same steady state), are closely connected with some further properties of the asymptotic behavior of the processes, like the asymptotic variance of certain functionals (performance measures and cost functions) of the network processes, or the speed of convergence to stationarity via comparison of the spectral gap.

Given a prescribed network in equilibrium, our expectation is, that if we perturb the routing process (which governs the movements of the customers after being served at any node) in a way that makes it more dependent in a specified way, than the joint queue length process after such a perturbation will be more dependent in some (possibly differently) specified way.

We concentrate especially on two ways in which the routing process is perturbed. The first way is by making routing more chaotic, which is borrowed from statistical mechanics. There exists a well-established method to express more or less *chaotic behavior of a random walker*, if his itinerary is governed by doubly stochastic routing matrices, see Alberti and Uhlmann [2]. We shall prove that if the routing is becoming more chaotic in this sense then the joint queue length process will show less internal dependency.

While the perturbation of the routing in this case is not connected with any order (numbering) of the nodes of a network, the second way of perturbing the routing is connected to some preassigned order of the nodes, which is expressed by a graph structure. Assuming that routing of customers is compatible with this graph structure, we perturb it by shifting probability mass in the routing kernel along paths that are determined by the graph. We shall prove that if we shift some masses in a way that routing becomes more positive dependent then internal dependence of the joint queue length process will increase.

We denote the Kronecker-Delta by $\delta_{i,j} = \begin{cases} 1 & \text{if} \quad i = j \\ 0 & \text{if} \quad i \neq j \end{cases}$, and for any real valued vector $\xi = (\xi_i : 0 \leq i \leq J)$ we define the diagonal matrix with entries from ξ by

$$diag(\xi) = (\delta_{i,j} \cdot \xi_i : 0 \leq i, j \leq J).$$

For $k = 1, \dots, J$, the kth J-dimensional unit (row) vector is $e_k := (\delta_{jk} : j = 1, \dots, J)$.

For $\alpha = (\alpha_1,\dots,\alpha_J) \in \mathbb{R}^J$ the rank statistic $\mathcal{R}(\alpha) = (\mathcal{R}_1(\alpha),\dots,\mathcal{R}_J(\alpha)) \in \mathbb{N}^J$ is defined by the enumeration of the indices of α in the decreasing order of their associated $\alpha_{(.)}-$values, i.e.

$$\alpha_{\mathcal{R}_i(\alpha)} \geq \alpha_{\mathcal{R}_{i+1}(\alpha)} \quad i = 1,\dots,J-1,$$

and ties are resolved according to the natural order of the indices.
The vector $A\mathcal{R}(\alpha) = (A\mathcal{R}_1(\alpha),\dots,A\mathcal{R}_J(\alpha)) \in \mathbb{N}^J$ of antiranks of α is defined by $A\mathcal{R}_j(\alpha) = \mathcal{R}_{J+1-j}(\alpha)$, and so yields an enumeration of the indices of α in the increasing order of their associated $\alpha_{(.)}-$values.

8.4.1 Correlation inequalities via generators

For a queue length network process $\tilde{\mathbf{X}}$ with generator $Q^{\tilde{\mathbf{X}}}$ and stationary distribution $\tilde{\pi}^J$ we are interested in *one step* correlation expressions.

$$\langle f, Q^{\tilde{\mathbf{X}}} g\rangle_{\tilde{\pi}^J} \tag{8.10}$$

If $f = g$, then (8.10) is (the negative of) a quadratic form, because $-Q^{\tilde{\mathbf{X}}}$ is positive definite. (8.10) occurs in the definition of Cheeger's constant which is helpful to bound the second largest eigenvalue of $Q^{\tilde{\mathbf{X}}}$ (because division of (8.10) by $\langle f, f\rangle_{\tilde{\pi}^J}$ yields Rayleigh quotients), which essentially governs the (L_2) speed of convergence of $\tilde{\mathbf{X}}$ to its equilibrium.

(8.10) can be utilized to determine the asymptotic variance of some selected cost or performance measures associated with Markovian processes (network processes) and to compare the asymptotic variances of both such related processes.

In a natural way, the one step correlations occur when comparing the dependence structure of $\tilde{\mathbf{X}}$ with that of a related process $\tilde{\mathbf{X}}'$, which has the same stationary distribution $\tilde{\pi}^J$, where we evaluate

$$\langle f, Q^{\tilde{\mathbf{X}}} g\rangle_{\tilde{\pi}^J} - \langle f, Q^{\tilde{\mathbf{X}}'} g\rangle_{\tilde{\pi}^J}, \tag{8.11}$$

see e.g. (iv) and (v) in Theorem 8.4.15 below.

Because we are dealing with processes having bounded generators, properties connected with (8.10) can be turned into properties of

$$\langle f, I + \varepsilon Q^{\tilde{\mathbf{X}}} g\rangle_{\tilde{\pi}^J} = E_{\tilde{\pi}^J}(f(\tilde{X}(0))g(\tilde{X}(\tau))), \tag{8.12}$$

where I is the identity operator, and $\varepsilon > 0$ is sufficiently small such that $I + \varepsilon Q^{\tilde{\mathbf{X}}}$ is a stochastic matrix, and $\tau \sim \exp(\varepsilon)$ (exponentially distributed). This enables us to apply some known discrete time methods to characterize properties of continuous time processes in the range of problems sketched above.

We begin with expressions which connect the differences (8.11) of covariances for related network processes with some covariances for the corresponding routing matrices.

Property 8.4.1 *Suppose* $\tilde{\mathbf{X}}$ *is an ergodic Jackson network process with a routing matrix* $\tilde{R}$ *and* $\tilde{\mathbf{X}}'$ *is a Jackson network process having the same arrival and service intensities but having a routing matrix* $\tilde{R}' = [\tilde{r}'_{ij}]$*, such that the* extended traffic solutions $\tilde{\eta}$ *of the traffic equation for* $\tilde{R}$ *and for* $\tilde{R}'$ *coincide. Then for all real functions* f,g

$$\langle f, Q^{\tilde{\mathbf{X}}} g\rangle_{\tilde{\pi}^J} - \langle f, Q^{\tilde{\mathbf{X}}'} g\rangle_{\tilde{\pi}^J} = \frac{\lambda}{\xi_0} E_{\tilde{\pi}^J}\left(tr\left(W^{g,f}(\tilde{X}(t))\cdot diag(\xi)\cdot(\tilde{R}-\tilde{R}')\right)\right),$$

where ξ *is the probability solution of the extended traffic equation* (8.3)*,* $e_0 = (0,\ldots,0)$*, and*

$$W^{g,f}(\mathbf{n}) = [g(\mathbf{n}+e_i)f(\mathbf{n}+e_j)]_{i,j=0,1,\ldots,J}.$$

Property 8.4.2 *Suppose* $\mathbf{X}$ *is an ergodic Gordon-Newell network process with a routing matrix* R *and* $\mathbf{X}'$ *is a Gordon-Newell network process having the same service intensities but having a routing matrix* $R' = [r'_{ij}]$ *such that the* stochastic traffic solutions η *of the traffic equation for* R *and for* R' *coincide. Then for all real functions* f,g

$$\begin{aligned}&\langle f, Q^{\mathbf{X}} g\rangle_{\pi^{(N,J)}} - \langle f, Q^{\mathbf{X}'} g\rangle_{\pi^{(N,J)}}\\ &= \frac{G(N-1,J)}{G(N,J)} E_{\pi^{(N-1,J)}}\left(tr\left(W^{g,f}(X(t))\cdot diag(\eta)\cdot(R-R')\right)\right),\end{aligned}$$

where η *is the probability solution of the traffic equation* (8.1)*,* $e_0 = (0,\ldots,0)$*, and*

$$W^{g,f}(\mathbf{n}) = [g(\mathbf{n}+e_i)f(\mathbf{n}+e_j)]_{i,j=1,\ldots,J}.$$

We shall reformulate the results of these properties in a form which is of independent interest, because it immediately relates our results to methods utilized in optimizing MCMC simulation. Introducing for convenience the notation $H^f(\mathbf{n},i) := f(\mathbf{n}+e_i)$ which in our framework occurs as $H^f(X(t),i) := f(X(t)+e_i)$ (and similarly for g), we obtain

Corollary 8.4.3 **(a)** *For Jackson network processes* $\tilde{\mathbf{X}}, \tilde{\mathbf{X}}'$ *as in Property 8.4.1, with* ξ *the probability solution of the extended traffic equation* (8.3)*, we have*

$$\langle f, Q^{\tilde{\mathbf{X}}} g\rangle_{\tilde{\pi}^J} - \langle f, Q^{\tilde{\mathbf{X}}'} g\rangle_{\tilde{\pi}^J} = \frac{\lambda}{\xi_0} E_{\tilde{\pi}^J}\langle H^f(\tilde{X}(t),\cdot), (\tilde{R}-\tilde{R}')H^g(\tilde{X}(t),\cdot)\rangle_{\xi} \quad (8.13)$$

(b) *For Gordon-Newell network processes* $\mathbf{X}, \mathbf{X}'$ *as in Proposition 8.4.2, with* η *the probability solution of the traffic equation, we have*

$$\langle f, Q^{\mathbf{X}} g\rangle_{\pi^{(N,J)}} - \langle f, Q^{\mathbf{X}'} g\rangle_{\pi^{(N,J)}} \tag{8.14}$$
$$= \frac{G(N-1,J)}{G(N,J)} E_{\pi^{(N-1,J)}} \langle H^f(X(t),\cdot), \big(R-R'\big) H^g(X(t),\cdot)\rangle_{\eta}$$

There are several appealing interpretations of the formulas (8.13) and (8.14) which will guide some of our forthcoming arguments. We discuss the closed network case (8.14).

The inner product

$$\langle H^f(X(t),\cdot), \big(R-R'\big) H^g(X(t),\cdot)\rangle_{\eta}$$

can be evaluated path-wise for any ω, and whenever, e.g., the difference $R-R'$ is positive definite, the integral $E_{\pi^{I-1,J}}(\cdot)$ (across Ω) is over non negative functions. Recalling that η is invariant for R and R', we obtain

$$\langle H^f(X(t),\cdot), \big(R-R'\big) H^g(X(t),\cdot)\rangle_{\eta} =$$
$$E_{\eta}\left(H^f(X(t),V_0)\cdot H^g(X(t),V_1)\right) - E_{\eta}\left(H^f(X(t),V_0')\cdot H^g(X(t),V_1')\right),$$

where $V = (V_n : n \in \mathbb{N})$, and $V' = (V_n' : n \in \mathbb{N})$ are Markov (routing) chains with the common stationary distribution η, and with two different transition matrices R, R'. If we consider formally a network process $\mathbf{X}$, and Markov chains V, resp. V' that are independent of $\mathbf{X}$, we get

$$\langle f, Q^{\mathbf{X}} g)_{\pi^{(N,J)}} - (f, Q^{\mathbf{X}'} g\rangle_{\pi^{(N,J)}} =$$
$$\frac{G(N-1,J)}{G(N,J)}\cdot$$
$$\cdot\left(E_{\pi^{(N-1,J)}}\, E_{\eta}\left(H^f(X(t),V_0)\cdot H^g(X(t),V_1)\right) - E_{\pi^{(N-1,J)}}\, E_{\eta}\left(H^f(X(t),V_0')\cdot H^g(X(t),V_1')\right)\right) =$$
$$= \frac{G(N-1,J)}{G(N,J)}\cdot$$
$$\cdot\left(E_{\eta}\, E_{\pi^{(N-1,J)}}\left(H^f(X(t),V_0)\cdot H^g(X(t),V_1)\right) - E_{\eta}\, E_{\pi^{(N-1,J)}}\left(H^f(X(t),V_0')\cdot H^g(X(t),V_1')\right)\right),$$

the latter equality by the Fubini theorem.

Corollary 8.4.3 points out the relevance of the following orderings for transition matrices which are well known in the theory of optimal selection of transition kernels for MCMC simulation. In our investigations these orders will be utilized to compare routing processes via their transition matrices.

Definition 8.4.4 *Let $R = [r_{ij}]$ and $R' = [r_{ij}']$ be transition matrices on a finite set $\mathbb{E}$ such that $\eta R = \eta R' = \eta$.*

We say that R' is smaller than R in the positive definite order , $R' \prec_{pd} R$, if $R - R'$ is positive definite on $L_2(\mathbb{E}, \eta)$.

We say that R' is smaller than R in the Peskun order, $R' \prec_{P} R$, if for all $j, i \in \mathbb{E}$ with $i \neq j$, it holds $r_{ji}' \leq r_{ji}$, see Peskun [68].

Peskun used the latter order to compare reversible transition matrices having the same stationary distribution, and to compare their asymptotic variance. Tierney [89] proved that the main property used in the proof of Peskun, namely that $R \prec_P R'$ implies $R' \prec_{pd} R$, holds without reversibility assumptions.

Comparison of asymptotic variance

Peskun and Tierney derived comparison theorems for the asymptotic variance of Markov chains for application to optimal selection of MCMC transition kernels in discrete time. These asymptotic variances occur as the variance parameters in the limiting distribution in the central limit theorem for MCMC estimators.

In the setting of queueing networks, performance measures of interest usually are of the form $\pi(f) = E_{\tilde{\pi}^J}(f(\tilde{X}(t)))$. The value of them can be estimated as a time average, justified by the ergodic theorem for Markov processes, i.e. in the discrete time we have for large n

$$E_{\tilde{\pi}^J}(f(\tilde{X}(t))) \sim \frac{1}{n}\sum_{k=1}^{n} f(X_k).$$

Under some regularity conditions on a homogeneous Markov chain with a transition kernel K, there exists CLT of the form (weak convergence $\equiv \xrightarrow{D}$)

$$\sqrt{n}(\frac{1}{n}\sum_{k=1}^{n} f(X_k) - E_{\tilde{\pi}^J}(f(\tilde{X}(t)))) \xrightarrow{D} N(0, v(f,K)),$$

where the asymptotic variance is

$$v(f,K) = \langle f, f\rangle_{\tilde{\pi}^J} - \pi(f) + 2\sum_{k=1}^{\infty} \langle f, K^k f\rangle_{\tilde{\pi}^J}. \tag{8.15}$$

To arrange a discrete time framework for our network processes $\tilde{\mathbf{X}}$ we consider Markov chains with transition matrices of the form

$$K = I + \varepsilon Q^{\tilde{\mathbf{X}}}$$

(with $\varepsilon > 0$ sufficiently small) that occur in the compound Poisson representation of the transition probabilities of the network processes.

The next properties show that perturbing the routing in network can result in a larger asymptotic variance for the imbedded chain.

Property 8.4.5 **(a)** *Consider two ergodic Jackson networks with the same arrival and service intensities, and with the stationary queue length processes* $\tilde{\mathbf{X}}$ *and* $\tilde{\mathbf{X}}'$*. Assume that the corresponding* extended *routing matrices* $\tilde{R}$ *and* $\tilde{R}'$ *are reversible with respect to* ξ*.*

If $\tilde{R}' \prec_P \tilde{R}$ *then for any real function* f *we have*

$$v(f, I + \varepsilon Q^{\tilde{\mathbf{X}}'}) \geq v(f, I + \varepsilon Q^{\tilde{\mathbf{X}}}). \tag{8.16}$$

(b) *Consider two ergodic Gordon-Newell networks with the same service intensities, and with the stationary queue length processes* $\mathbf{X}$ *and* $\mathbf{X}'$. *Assume that the corresponding routing matrices* R *and* R' *are reversible with respect to* η.

If $R' \prec_P R$, *then for any real function* f *we have*

$$v(f, I + \varepsilon Q^{\mathbf{X}'}) \geq v(f, I + \varepsilon Q^{\mathbf{X}}). \tag{8.17}$$

Comparison of spectral gaps

Let $\mathbf{X}$ be a continuous time homogeneous ergodic Markov process with stationary probability π, and generator $Q^{\mathbf{X}}$. The spectral gap of $\mathbf{X}$, resp. $Q^{\mathbf{X}}$ is

$$Gap(Q^{\mathbf{X}}) = \inf\{\langle f, -Q^{\mathbf{X}} f\rangle_\pi : f \in L_2(\mathbb{E}, \pi), \pi(f) = 0,\ \langle f, f\rangle\pi = 1\}. \tag{8.18}$$

The spectral gap of $\mathbf{X}$ determines for $X(t)$ the distance to equilibrium π in $L_2(\mathbb{E}, \pi)$-norm $\|\cdot\|_\pi$: $Gap(Q^{\mathbf{X}})$ is the largest number Δ such that for the transition semigroup $P = (P_t : t \geq 0)$ of $\mathbf{X}$ it holds

$$\|P_t f - \pi(f)\|_\pi \leq e^{-\Delta t}\|f - \pi(f)\|_\pi \quad \forall f \in L_2(\mathbb{E}, \pi).$$

For Gordon-Newell networks their spectral gap is always greater than zero, while for Jackson networks the situation is more delicate: zero gap and non zero gap can occur. Iscoe and McDonald [45], [46], and Lorek [56] proved, under some natural assumptions, necessary and sufficient conditions for the existence of the non-zero spectral gap of Jackson networks. The case of positive gap is proved by using an auxiliary vector of independent birth-death processes, used to bound the gap away from zero.

It is interesting that for some classes of Jackson networks it is possible to strictly bound the gap of the queue length network process $\tilde{\mathbf{X}}$ from below by the gap of some multidimensional birth-death process, which will play in the next statement the role of the network process $\tilde{\mathbf{X}}'$. Because we focus on the intuitive, but rather strong Peskun ordering of the routing matrices, we need some additional assumptions on the routing. The assumption constitutes a detailed balance which determines an additional internal structure of a Markov chain and its global balance equation (= equilibrium equation). Such detailed balance equations are prevalent in many networks with (nearly) product form steady states, and often open a way to solve the global balance equation for the steady state. (8.19) equalizes the routing flow from any node into the (inner) network to the flow out of the (inner) network to that node.

Property 8.4.6 *Consider an ergodic Jackson network process* $\tilde{\mathbf{X}}$ *with* $\lambda_i > 0$, *for* $i = 1, \ldots, J$. *Assume that the corresponding extended routing matrix* $\tilde{R} = [\tilde{r}_{ij}]_{i,j=0,1,\ldots,J}$ *has strict positive departure probabilities* $\tilde{r}_{i0} > 0$ *from every node* $i = 1, \ldots, J$.

Assume further that the routing of $\tilde{\mathbf{X}}$ *fulfills the following overall balance for all network nodes with respect to the solution* $\tilde{\eta}_i, i = 1, \ldots, J$, *of the traffic equation* (8.3)*:*

$$\tilde{\eta}_j \sum_{i=1}^{J} \tilde{r}_{j,i} = \sum_{i=1}^{J} \tilde{\eta}_i \tilde{r}_{i,j}, \quad \forall j = 1, \ldots, J. \tag{8.19}$$

Then for the vector valued process $\mathbf{X}^*$ *consisting of the independent birth-death processes, for which the nodes have the same service intensities as the Jackson nodes and the external arrival rates* $\lambda_i^* = \lambda_i$ *it holds*

$$Gap(Q^{\tilde{\mathbf{X}}}) \geq Gap(Q^{\mathbf{X}^*}).$$

So, we can immediately conclude for some networks that $Gap(Q^{\tilde{\mathbf{X}}}) \geq Gap(Q^{\mathbf{X}^*})$ holds. A consequence which elaborates on the implication *Peskun yields positive definiteness* is, that if we perturb routing of customers in the networks by shifting mass from non diagonal entries to the diagonal (leaving the routing equilibrium fixed) and obtaining that way the Peskun order of routing, then the speed of convergence of the perturbed process can only decrease. This is just what in optimization of MCMC was intended, and Peskun gave conditions for this. Similarly we see

Property 8.4.7 **(a)** *Consider two ergodic Jackson networks with the same arrival and service intensities, and with the state processes* $\tilde{\mathbf{X}}$ *and* $\tilde{\mathbf{X}}'$. *Assume that for the* extended *routing matrices* $\tilde{R}$ *and* $\tilde{R}'$, *the stochastic solutions of the traffic equations coincide (being* ξ *). If* $\tilde{R} \prec_{pd} \tilde{R}'$ *then for any real function* f *we have*

$$\langle f, Q^{\tilde{\mathbf{X}}'} f \rangle_{\tilde{\pi}^J} \geq \langle f, Q^{\tilde{\mathbf{X}}} f \rangle_{\tilde{\pi}^J}, \quad and \quad Gap(Q^{\tilde{\mathbf{X}}'}) \leq Gap(Q^{\tilde{\mathbf{X}}}). \tag{8.20}$$

(b) *Consider ergodic Gordon-Newell networks with the same service intensities, and with the state processes* $\mathbf{X}$ *and* $\mathbf{X}'$. *Assume that for the corresponding routing matrices* R *and* R' *the stochastic solutions of the traffic equations coincide. If* $R \prec_{pd} R'$ *then for any real function* f *we have*

$$\langle f, Q^{\mathbf{X}'} f \rangle_{\pi^{(N,J)}} \geq \langle f, Q^{\mathbf{X}} f \rangle_{\pi^{(N,J)}}, \quad and \quad Gap(Q^{\mathbf{X}'}) \leq Gap(Q^{\mathbf{X}}). \tag{8.21}$$

Comparison of dependencies

The expression (8.10) for continuous time Markov processes is transformed via the embedded uniformization chain (8.12) to a covariance value and via (8.11) to a comparison statement for covariance functionals for two Markov processes (and their Poissonian embedded chains), i.e., with $\tau \sim \exp(\eta)$, we obtain

$$E_{\tilde{\pi}^J}(f(\tilde{X}_0)g(\tilde{X}_\tau)) = \langle f, (I + \eta Q^{\tilde{\mathbf{X}}})g \rangle_{\tilde{\pi}^J} \leq \langle f, (I + \eta Q^{\tilde{\mathbf{X}}'} \rangle g \rangle_{\tilde{\pi}^J} = E_{\tilde{\pi}^J}(f(\tilde{X}'_0)g(\tilde{X}'_\tau)).$$

A procedure of transforming this property into analogous statements for the continuous time evolution (over many time points) will need in general some additional monotonicity properties of the processes under consideration. It turns out that some form of stochastic monotonicity is in some cases a direct substitute for the strong reversibility assumption which is needed to prove Peskun's theorem.

8.4.2 Doubly stochastic routing

In this section the perturbation of a network process will be due to the fact that the routing of the customers will become more chaotic. In statistical physics there is a well-established method to express being more or less *chaotic* for a random walker, if his itinerary is governed by a doubly stochastic routing matrix. Alberti and Uhlmann provide an in-depth study of *Stochasticity and Partial Order* that elaborates on such methods [2]. Following their ideas, we shall consider (mainly) Gordon-Newell networks with doubly stochastic routing matrix.

Consider an arbitrary row $r(i) := (r_{ij} : j = 1,2,\dots,J)$ of the Gordon-Newell network's routing matrix R and a doubly stochastic matrix $T = [t_{ij}]_{i,j=1,\dots,J}$. Then the i-th row vector of the product $(R \cdot T)$ is smaller than $r(i)$ in the sense of the majorization ordering, see Marshall and Olkin [57], p.18. This means that the probability mass is more equally distributed in each row after multiplication. The routing scheme is then more equally distributed too. Nevertheless, the solution of the traffic equation for $R \cdot T$ and therefore the steady state of the network under the $R \cdot T$ regime is the same as under R, namely, the normalized solution of the traffic equation (8.1) is in both cases the uniform distribution on $\{1,2,\dots,J\}$.

A more chaotic routing leads to less internal dependencies over time of the individual routing chains of the customers and will therefore lead to less internal dependence over time of the joint queue length process. Let

$$\mathcal{L} = \{f : \mathbb{E}_N \to \mathbb{R}_+ : f(n_1,\dots,n_J) = a + \sum_{i=1}^{J} \alpha_i \cdot n_i, \alpha_i \in \mathbb{R}, i = 1,\dots,J, a \in \mathbb{R}_+\},$$

be the convex cone of nonnegative affine-linear functions on $\mathbb{E}_N$.

Theorem 8.4.8 (Linear service rates) *Consider two ergodic Gordon-Newell network processes with common stationary distribution $\pi^{(N,J)}$:* $\mathbf{X}$ *with a doubly stochastic routing matrix R and* $\mathbf{X}'$ *with the routing matrix $R' = [r'_{ij}] = R \cdot T$, for a doubly stochastic matrix T. All other parameters of the networks are assumed to be the same.*

Consider pairs of nonnegative affine-linear functions

$$f : \mathbb{E}_N \to \mathbb{R}_+ : f(n_1,\dots,n_J) = a + \sum_{i=1}^{J} \alpha_i \cdot n_i \in \mathcal{L}, \quad \textit{and}$$

$$g : \mathbb{E}_N \to \mathbb{R}_+ : g(n_1,\dots,n_J) = b + \sum_{i=1}^{J} \beta_i \cdot n_i \in \mathcal{L}, \quad \textit{with}$$

$$\mathcal{R}(\alpha_1,\dots,\alpha_J) = \mathcal{R}(\beta_1,\dots,\beta_J).$$

Then for all such pairs of functions with $f,g \in \mathcal{I}^*_+(\mathbb{N}^J) \cap \mathcal{L}$, and $f,g \in \mathcal{D}^*_+(\mathbb{N}^J) \cap \mathcal{L}$, holds

$$\langle f, Q^{\mathbf{X}'} g \rangle_{\pi^{(N,J)}} \le \langle f, Q^{\mathbf{X}} g \rangle_{\pi^{(N,J)}}.$$

In Theorem 8.4.8, for $f = g$, the rank condition is trivially fulfilled. This yields

Corollary 8.4.9 *Under the assumptions of Theorem 8.4.8, for all $f \in \mathcal{I}^*_+(\mathbb{N}^J) \cap \mathcal{L}$, and $f \in \mathcal{D}^*_+(\mathbb{N}^J) \cap \mathcal{L}$, it holds*

$$\langle f, Q^{\mathbf{X}'} f \rangle_{\pi^{(N,J)}} \leq \langle f, Q^{\mathbf{X}} f \rangle_{\pi^{(N,J)}}.$$

Note, that for $f = g$, $R(I - T)$ is nonnegative definite.

8.4.3 Robin-Hood transforms

If the node set J is equipped with a partial order, which is relevant for the customers' migration, then it is tempting to consider perturbations of the routing processes that are in line with this order. To be more precise: We have an up-down relation between the nodes and the question is how the steady state performance reacts on routing being more up, resp. down.

The construction of Corollary 2.1 and Example 3.1 in Daduna and Szekli [24], which is sometimes called ROBIN-HOOD TRANSFORM - since in a certain sense it equalizes the frequencies of the random walker to visit different nodes, yields a change of the routing pattern in networks. The construction is as follows:

Consider a homogeneous Markov chain (X_i) on a finite partially ordered state space $(\mathbb{E}, \prec)$ with a transition matrix $[p(i,j)]_{i,j \in \mathbb{E}}$, and the corresponding stationary distribution π.

Assume that for $a,b,c,d \in \mathbb{E}$, we have $a \prec c$ and $b \prec d$ such that $(a,d) \in \mathbb{E}^2$ and $(c,b) \in \mathbb{E}^2$ are not comparable with respect to the product order, and that $P^{(X_0,X_1)}(a,d) \geq \alpha, P^{(X_0,X_1)}(c,b) \geq \alpha$.

Construct the distribution $P^{(Y_0,Y_1)}$ of a random vector (Y_0,Y_1) from $P^{(X_0,X_1)}$ by

$$P^{(Y_0,Y_1)}(a,b) = P^{(X_0,X_1)}(a,b) + \alpha,\ P^{(Y_0,Y_1)}(c,d) = P^{(X_0,X_1)}(c,d) + \alpha, \text{ and}$$
$$P^{(Y_0,Y_1)}(a,d) = P^{(X_0,X_1)}(a,d) - \alpha,\ P^{(Y_0,Y_1)}(c,b) = P^{(X_0,X_1)}(c,b) - \alpha, \text{ and}$$
$$P^{(Y_0,Y_1)}(u,v) = P^{(X_0,X_1)}(u,v) \text{ for all other } (u,v) \in \mathbb{E}^2.$$

(This is the Robin-Hood transform.)

The one-dimensional marginals of both (X_0,X_1) and (Y_0,Y_1) are π and conditional distribution $P(Y_1 = w \mid Y_0 = v) =: q(v,w)$, for $v,w \in \mathbb{E}$, is obtained from $[p(i,j)]$ as follows:

$$\begin{aligned} q(a,d) &= p(a,d) - \frac{\alpha}{\pi(a)},\ q(c,b) = p(c,b) - \tfrac{\alpha}{\pi(c)}, \\ q(a,b) &= p(a,b) + \frac{\alpha}{\pi(a)},\ q(c,d) = p(c,d) + \tfrac{\alpha}{\pi(c)},\ q(u,v) = p(u,v) \quad \text{otherwise.} \end{aligned} \tag{8.22}$$

Consider now a homogeneous Markov chain (Y_i) with the so defined transition matrix q, and consider (X_i) and (Y_i) as routing chains of a network process, where (Y_i) is obtained from (X_i) by a perturbation through the Robin-Hood transformation.

Then according to Corollary 2.1 and Theorem 3.1 in Daduna and Szekli [24] the routing governed by (Y_i) is more concordant than the routing governed by (X_i).

Definition 8.4.10 *Let $(\mathbb{E},\prec)$ be a finite partially ordered set. The generalized partial sum order $\prec_*$ on $\mathbb{N}^{\mathbb{E}}$ is defined for $x=(x_i : i\in\mathbb{E}), y=(y_i : i\in\mathbb{E})\in\mathbb{N}^{\mathbb{E}}$ by*

$$x \prec_* y :\Longleftrightarrow \forall \text{ decreasing } K\subseteq\mathbb{E} \text{ holds } \sum_{k\in K} x_k \le \sum_{k\in K} y_k. \tag{8.23}$$

Consider now a Jackson network $\tilde{\mathbf{X}}$ where the node set $\mathrm{J}=\{1,\ldots,J\}$ is a partially ordered set $(\mathrm{J},\prec)$ and the customers flow in line with the directions prescribed by this partial order, i.e. for the routing matrix $\tilde{R}$ it holds (see Harris [33]):

$$\tilde{r}_{ij}>0 \Longrightarrow (i\prec j\vee j\prec i). \tag{8.24}$$

Then the Jackson network process $\tilde{\mathbf{X}}$ has the up-down property with respect to $\prec_*$, which means that for the generator $Q^{\tilde{\mathbf{X}}}$

$$q^{\tilde{\mathbf{X}}}(x,y)>0 \Longrightarrow (x\prec_* y\vee y\prec_* x). \tag{8.25}$$

Lemma 8.4.11 *Consider an ergodic Jackson network with an extended routing matrix $\tilde{R}$, and the corresponding queue length process $\tilde{\mathbf{X}}$. We assume that the node set $\mathrm{J}=\{1,\ldots,J\}$ is a partially ordered set. For some nodes $a,b,c,d\in\mathrm{J}$ (not necessarily distinct) let $a\prec c$ and $b\prec d$, and for some $\alpha>0$ let*

$$\tilde{r}_{ad}\ge\alpha/\xi_a \qquad \text{and} \qquad \tilde{r}_{cb}\ge\alpha/\xi_c. \tag{8.26}$$

Define a new Jackson network with its queue length process $\tilde{\mathbf{X}}'$ as follows: the nodes' structure, and the external arrival processes are the same as in the original network. The routing matrix $\tilde{R}'$ is computed by the Robin-Hood transformation (8.22) with the fixed a,b,c,d.

Consider a pair of comonotone functions f,g (either both increasing or both decreasing) such that for all $n\in\mathbb{N}^J$ it holds $(f(\mathbf{n}+\mathbf{e}_c)-f(\mathbf{n}+\mathbf{e}_a))\cdot(g(\mathbf{n}+\mathbf{e}_d)-g(\mathbf{n}+\mathbf{e}_b))\ge 0$. Then

$$\langle f,Q^{\tilde{\mathbf{X}}}g\rangle_{\tilde{\pi}^J}\le\langle f,Q^{\tilde{\mathbf{X}}'}g\rangle_{\tilde{\pi}^J}. \tag{8.27}$$

Immediately from this lemma we get

Theorem 8.4.12 *Consider an ergodic Jackson network with an extended routing matrix $\tilde{R}$, and with the corresponding queue length process $\tilde{\mathbf{X}}$. We assume that the node set is partially ordered $(\mathrm{J},\prec)$.*

Define a new Jackson network with its queue length process $\tilde{\mathbf{X}}'$ as follows: the nodes' structure, and the external arrival processes are the same as in the original network. The routing matrix $\tilde{R}'$ is computed by a sequence of $n\ge 1$ feasible Robin-Hood transformations according to (8.22) for a sequence of nodes.

Then for any pair of comonotone functions $f,g:\mathbb{N}^J\to\mathbb{R}_+$ with respect to the generalized partial sum order $\prec_*$ (either both increasing or both decreasing) it *holds*

$$\langle f, Q^{\tilde{\mathbf{X}}} g\rangle_{\tilde{\pi}^J} \leq \langle f, Q^{\tilde{\mathbf{X}}'} g\rangle_{\tilde{\pi}^J}.$$

8.4.4 Dependence orderings and monotonicity

We shall now generalize the concordance ordering.

Definition 8.4.13 **(a)** *Random elements* $\mathbf{X}, \mathbf{Y}$ *of* $\mathbb{E}^n$ *are called concordant stochastically ordered with respect to* $\mathcal{F}$ *(written as* $\mathbf{X} \prec^n_{\mathcal{F}-cc} \mathbf{Y}$ *or* $\mathbf{Y} \succ^n_{\mathcal{F}-cc} \mathbf{X}$*, often shortly:* $\mathbf{X} \prec_{\mathcal{F}-cc} \mathbf{Y}$*, resp.,* $\mathbf{Y} \succ_{\mathcal{F}-cc} \mathbf{X}$*,) if*

$$E\left[\prod_{i=1}^{n} f_i(X_i)\right] \leq E\left[\prod_{i=1}^{n} f_i(Y_i)\right], \tag{8.28}$$

for all $f_i \in \mathcal{I}^*_+(\mathbb{E}) \cap \mathcal{F}$ *and for all* $f_i \in \mathcal{D}^*_+(\mathbb{E}) \cap \mathcal{F}, i = 1, \ldots, n$.

(b) *Let* $T \subseteq \mathbb{R}$ *be an index set for stochastic processes* $\mathbf{X} = (X_t : t \in T)$ *and* $\mathbf{Y} = (Y_t : t \in T)$*,* $X_t, Y_t : \Omega \to \mathbb{E}$*,* $t \in T$*. We say that* $\mathbf{X}$ *and* $\mathbf{Y}$ *are concordant stochastically ordered with respect to a class* $\mathcal{F}$ *of functions on* $\mathbb{E}$ *(and write* $\mathbf{X} \prec_{\mathcal{F}-cc} \mathbf{Y}$*) if for all* $n \geq 2$ *and all* $t_1 < t_2 < \ldots t_n$*, we have on* $\mathbb{E}^n$

$$(X_{t_1}, \ldots, X_{t_n}) \prec_{\mathcal{F}-cc} (Y_{t_1}, \ldots, Y_{t_n}).$$

The setting of **(b)** will be applied to Markovian processes.

Taking in **(a)** for $\mathcal{F}$ the space of all measurable functions $\mathcal{M}$ on $\mathbb{E}$ we obtain the usual concordance ordering as in Daduna and Szekli [24]. It is easy to see that the two-dimensional marginals of the Markov chains related by the Robin-Hood construction in (8.22) fulfill

$$(X_0, X_1) \leq_{\mathcal{M}-cc} (Y_0, Y_1).$$

For example, if $\mathcal{F}$ contains the indicator functions of point-generated increasing and decreasing sets, $\{i\}^{\uparrow} = \{j \in E : i \prec j\}$ and $\{i\}^{\downarrow} = \{j \in E : j \prec i\}$, for concordant stochastically ordered processes $\mathbf{X}$ and $\mathbf{Y}$ (with respect to $\mathcal{F}$) we can compare the probability of extreme events like

$$P(\inf(X_{t_1}, \ldots, X_{t_n}) \succ t) \leq P(\inf(Y_{t_1}, \ldots, Y_{t_n}) \succ t),$$

and

$$P(\sup(X_{t_1}, \ldots, X_{t_n}) \prec s) \leq P(\sup(Y_{t_1}, \ldots, Y_{t_n}) \prec s),$$

for fixed t and s. We mention, that in most cases $\mathcal{F}$ will be a convex cone of functions which is often additionally closed under the point-wise convergence.

Discrete time.

Let $\mathbf{X} = (X_t : t \in \mathbb{Z})$ and $\mathbf{Y} = (Y_t : t \in \mathbb{Z}), X_t, Y_t : \Omega \to \mathbb{E}$, be discrete time, stationary, homogeneous Markov processes. Assume that π is the corresponding unique invariant (stationary) one–dimensional marginal distribution, the same for both $\mathbf{X}$ and $\mathbf{Y}$, and denote the 1–step transition kernels for $\mathbf{X}$ and $\mathbf{Y}$, by $K^X : \mathbb{E} \times \mathcal{E} \to [0,1]$, and $K^Y : \mathbb{E} \times \mathcal{E} \to [0,1]$, respectively. Denote the respective transition kernels for the time reversed processes $\overleftarrow{\mathbf{X}}$, $\overleftarrow{\mathbf{Y}}$ by $\overleftarrow{K^X}$, $\overleftarrow{K^Y}$. We say that a stochastic kernel $K : \mathbb{E} \times \mathcal{E} \to [0,1]$ is $\mathcal{F}$-monotone if $\int f(x)K(s,dx) \in \mathcal{I}^*_+(\mathbb{E}) \cap \mathcal{F}$ for each $f \in \mathcal{I}^*_+(\mathbb{E}) \cap \mathcal{F}$.

The following property proved to be useful in comparing some second order properties of Markov processes, see Hu and Pan [34], Daduna and Szekli [23], Baeuerle and Rolski [5], Daduna and Szekli [24]. It will be convenient to impose this condition here as well. A pair $\mathbf{X}$ and $\mathbf{Y}$ of discrete time Markov processes having the same invariant probability measure fulfils

$\mathcal{F}$-symmetric monotonicity if : *Either K^Y and $\overleftarrow{K^X}$ are $\mathcal{F}$-monotone, or K^X and $\overleftarrow{K^Y}$ are $\mathcal{F}$- monotone.*

The following theorem is an analog of Theorem 3.1 in Daduna and Szekli [24].

Theorem 8.4.14 (concordance ordering under $\mathcal{F}$- symmetric monotonicity) *For two stationary Markov processes $\mathbf{X}, \mathbf{Y}$ defined above, having the common unique invariant distribution π, and fulfilling $\mathcal{F}$-symmetric monotonicity, the following relations are equivalent*

(i) $\mathbf{X} \prec_{\mathcal{F}-cc} \mathbf{Y}$
(ii) $(X_0, X_1) \prec^2_{\mathcal{F}-cc} (Y_0, Y_1)$
(iii) $\langle f, K^X g\rangle_\pi \leq \langle f, K^Y g\rangle_\pi$ *for all* $f, g \in \mathcal{I}^*_+(\mathbb{E}) \cap \mathcal{F}$, *and for all* $f, g \in \mathcal{D}^*_+(\mathbb{E}) \cap \mathcal{F}$
(iv) $\langle f, \overleftarrow{K^X} g\rangle_\pi \leq \langle f, \overleftarrow{K^Y} g\rangle_\pi$ *for all* $f, g \in \mathcal{I}^*_+(\mathbb{E}) \cap \mathcal{F}$, *and for all* $f, g \in \mathcal{D}^*_+(\mathbb{E}) \cap \mathcal{F}$

Continuous time.

Let $\mathbf{X} = (X_t : t \in \mathbb{R})$ and $\mathbf{Y} = (Y_t : t \in \mathbb{R}), X_t, Y_t : \Omega \to \mathbb{E}$, be stationary homogeneous Markov processes with countable state spaces. Denote the corresponding families of transition kernels of $\mathbf{X}$, and $\mathbf{Y}$, by $\mathbb{K}^X = (K^X_t : \mathbb{E} \times \mathcal{E} \to [0,1] : t \geq 0)$, and $\mathbb{K}^Y = (K^Y_t : \mathbb{E} \times \mathcal{E} \to [0,1] : t \geq 0)$, respectively, and the respective transition kernels for the stationary time reversed processes $\overleftarrow{\mathbf{X}}$, $\overleftarrow{\mathbf{Y}}$ by $\overleftarrow{\mathbb{K}^X} = (\overleftarrow{K^X_t} : \mathbb{E} \times \mathcal{E} \to [0,1] : t \geq 0)$, and $\overleftarrow{\mathbb{K}^Y} = (\overleftarrow{K^Y_t} : \mathbb{E} \times \mathcal{E} \to [0,1] : t \geq 0)$, respectively. Assume that π is the corresponding invariant distribution, common for both $\mathbb{K}^X$ and $\mathbb{K}^Y$, that is $\int K^X_t(x,dy)\pi(dx) = \int K^Y_t(x,dy)\pi(dx) = \pi(dy)$, for all $t > 0$.

For the time reversed processes we use the corresponding notation $\overleftarrow{Q^X}$ and $\overleftarrow{Q^Y}$. We say that $\mathbb{K}^X = (K^X_t : \mathbb{E} \times \mathcal{E} \to [0,1] : t \geq 0)$ is $\mathcal{F}$-time monotone if for each $t \geq 0$, K^X_t is $\mathcal{F}$- monotone.

Analogously to the discrete case, we define: A pair **X** and **Y** of continuous time Markov processes having the same invariant probability measure fulfills

$\mathcal{F}$-time symmetric monotonicity if : *Either $\mathbb{K}^Y$ and $\overleftarrow{\mathbb{K}}^X$ are $\mathcal{F}$-time monotone, or $\mathbb{K}^X$ and $\overleftarrow{\mathbb{K}}^Y$ are $\mathcal{F}$- time monotone.*

Using similar arguments as in Theorem 3.3 in Daduna and Szekli [24] we have

Theorem 8.4.15 *Suppose that $\mathbb{E}$ is countable and the above defined stationary chains **X** and **Y** have bounded intensity matrices Q^X and Q^Y, respectively. Then under $\mathcal{F}$-time symmetric monotonicity the following properties are equivalent*

(i) $\mathbf{X} \prec_{\mathcal{F}-cc} \mathbf{Y}$
(ii) $(X_0, X_t) \prec^2_{\mathcal{F}-cc} (Y_0, Y_t) \quad \forall t > 0,$
(iii) $\langle f, T_t^X g\rangle_\pi \le \langle f, T_t^Y g\rangle_\pi$ *for all* $f, g \in \mathcal{I}^*_+(\mathbb{E}) \cap \mathcal{F}$, *and for all* $f, g \in \mathcal{D}^*_+(\mathbb{E}) \cap \mathcal{F}, \forall t > 0$
(iv) $\langle f, Q^X g\rangle_\pi \le \langle f, Q^Y g\rangle_\pi$ *for all* $f, g \in \mathcal{I}^*_+(\mathbb{E}) \cap \mathcal{F}$, *and for all* $f, g \in \mathcal{D}^*_+(\mathbb{E}) \cap \mathcal{F}$
(v) $\langle f, \overleftarrow{Q}^X g\rangle_\pi \le \langle f, \overleftarrow{Q}^Y g\rangle_\pi$ *for all* $f, g \in \mathcal{I}^*_+(\mathbb{E}) \cap \mathcal{F}$, *and for all* $f, g \in \mathcal{D}^*_+(\mathbb{E}) \cap \mathcal{F}$

Reducing the class of functions from $\mathcal{M}$ to some smaller class $\mathcal{F}$ makes this theorem much more versatile for applications, as we shall demonstrate below.

From Theorem 8.4.15, we conclude that problem of comparing correlations for stochastic network processes in continuous time is an interplay of two tasks:
• proving monotonicity, the form of which we identified as $\mathcal{F}$- time symmetric monotonicity, and
• additionally proving generator inequalities.

Generator inequalities have been presented in the previous paragraphs. We shall continue with presenting the concept of *time symmetric monotonicity* for network processes.

From a recent literature on dependence structure of Markovian processes with one dimensional (linearly ordered) discrete state spaces it is visible that *$\mathcal{F}$-time symmetric monotonicity* (in continuous time) and *$\mathcal{F}$ symmetric monotonicity* (in discrete time) play an important role, see e.g., Hu and Pan [34]. This property occurred independently in the literature several times, see e.g., Baeuerle and Rolski [5], Daduna and Szekli [23][Lemma 3.2].

So - in general, we cannot hope to dispense from these assumptions when proving dependence properties in a more complex network setting. Nevertheless, the necessity of these assumptions is still an unsolved problem. Some counterexamples, where dependence structures of Markovian processes over a finite time horizon are proved without *$\mathcal{F}$ symmetric monotonicity*, are provided in Daduna and Szekli [24][Section 3.3].

On the other hand, a need for some monotonicity is emphasized further by the related theory of association in time for Markov processes, which relies on the strong stochastic monotonicity of these processes, see for a review Liggett [50][chapter II], and Daduna and Szekli [23].

For stochastic networks, which are in general not reversible, the property of *time symmetric monotonicity* seems to ba a natural property: Every Jackson network process $\mathbf{X}$ with service rates that are at all nodes nondecreasing functions of the local queue length [Daduna and Szekli [23],Corollary 4.1] is stochastically monotone with respect to stochastic ordering on the set of all probability measures on $(\mathbb{N}^J, \leq)$. Because the time reversed process of a Jackson network process is equal in distribution to a process of a suitably defined Jackson network with the same properties for the service rates, any pair of Jackson network processes with the same stationary distribution fulfills $\mathcal{F}$*-time symmetric monotonicity*, where $\mathcal{F} = \mathcal{I}^*(\mathbb{N}^J, \leq)$.

We only mention that by a similar observation $\mathcal{F}$*-time symmetric monotonicity* holds for Gordon-Newell networks.

In many papers $\mathcal{F}$ is the class of all (bounded) increasing functions with respect to the natural linear order. The weaker concept of $\mathcal{F}$*-(time) symmetric monotonicity* for smaller classes of functions seems to be natural in the context of the theory of integral orders, see Mueller and Stoyan [65] or Li and Shaked [54]. However we shall need a *closure property*, which will guarantee that $\mathcal{F}$-functions are transformed into $\mathcal{F}$-functions, or at least into the maximal generator of the respective order, Mueller and Stoyan [65][Definition 2.3.3] or Li and Shaked [54] (Definition 3.2).

The balance between having a small class of $\mathcal{F}$-functions and the necessity of obtaining such a closure property is demonstrated next. The first example is in the spirit of the classical Gordon-Newell networks but with a smaller set $\mathcal{F}$. Recall that $\mathcal{L}$ is the set of nonnegative affine-linear functions on $\mathbb{E}_N$.

Property 8.4.16 (Linear service rates) *Consider two Gordon-Newell network processes $\mathbf{X}, \mathbf{X}'$ on $\mathbb{E}_N \subseteq \mathbb{N}^J$, equipped with the coordinate-wise order $\leq$, both with the corresponding stationary distribution $\pi^{N,J}$. Assume that the service rates in both networks at all nodes are linear functions of the local queue lengths.*

Then the pair $\mathbf{X}, \mathbf{X}'$ of Gordon-Newell network processes is $\mathcal{L}$-time symmetric monotone.

Property 8.4.17 (Generalized tandem network) *Consider an open tandem network process $\tilde{\mathbf{X}}$ on the state space $\mathbb{N}^J$ equipped with the partial sum order $\leq_*$ with stationary distribution $\tilde{\pi}^J$. The routing for $\tilde{\mathbf{X}}$ is* linear *as follows:*

- *customers enter the network only through node* 1*:* $\lambda_1 > 0, \lambda_j = 0, j = 2, \ldots, J,$
- *customers depart from the network only from node J:* $\tilde{r}_{J0} > 0, r_{j0} = 0, j = 1, \ldots, J-1,$
- *customers move only stepwise:* $\tilde{r}_{j(j+1)} > 0, j = 1, \ldots, J-1,$ *and* $\tilde{r}_{j(j-1)}) \geq 0, j = 2, \ldots, J,$
  and $\tilde{r}_{jj} \geq 0, j = 1, \ldots, J,$ *and* $\tilde{r}_{ji} = 0$ *in any other case.*

Let $\tilde{\mathbf{X}}'$ be another generalized tandem network process with the same stationary distribution $\tilde{\pi}^J$, and with its routing subject to the same restriction as described for $\tilde{\mathbf{X}}$.

Assume that the arrival rates and the (nondecreasing) service rates in both networks are equal and bounded.

Then the pair $\tilde{\mathbf{X}}, \tilde{\mathbf{X}}'$ is $\mathcal{I}^*(\mathbb{R}^J, \leq_*) \cup \mathcal{D}^*(\mathbb{R}^J, \leq_*)$-time symmetric monotone.

Property 8.4.18 (Functions of the total population size) *Consider two Jackson networks with equal linear service rates, which have the same stationary distribution. Assume further that inside both networks the effective departure rates from all nodes are the same, i.e., $\mu_j \cdot r_{j0}$ is constant for all $j = 1, \dots, J$, (and therefore > 0). Let*

$$\mathcal{F} = \{f : \mathbb{N}^J \to \mathbb{R}_+ : f(n_1, \dots, n_J) = f^*(n_1 + \dots + n_J) \textit{ for some } f^* : \mathbb{R} \to \mathbb{R}_+\}$$

be the set of real valued functions on $\mathbb{N}^J$, which depend on the sum of the arguments only.
Then the state processes in these networks constitute a $\mathcal{F}-$time symmetric monotone pair.

Let $\rho = (\rho_1, \dots, \rho_J)$ be a permutation of $\{1, 2, \dots, J\}$ which will serve as a rank vector for the linear factors of functions in

$$\begin{aligned}\mathcal{L}(\rho) = \{f : S(I,J) \to \mathbb{R}_+ : f(n_1, \dots, n_J) \qquad\qquad\qquad\qquad\qquad\qquad (8.29)\\ = a + \sum_{i=1}^{J} \alpha_i \cdot n_i, \alpha_i \in \mathbb{R}, i = 1, \dots, J, a \in \mathbb{R}_+, \mathcal{R}(\alpha_1, \dots, \alpha_J) = \rho\} \subseteq \mathcal{L}.\end{aligned}$$

Theorem 8.4.19 *Consider two ergodic Gordon-Newell network processes with common stationary distribution $\pi^{(N,J)}$: $\mathbf{X}$ with a doubly stochastic routing matrix $R = [r_{ij}]$ and $\mathbf{X}'$ with the routing matrix $R' = R \cdot T$, for a doubly stochastic matrix $T = [t_{ij} : i, j = 1, \dots, J]$. The service rates $\mu_j(n_j) = \mu_j \cdot n_j$ are in both networks the same.*

Let $A\mathcal{R}(\mu) = \rho = (\rho_1, \dots, \rho_J)$ denote the antirank vector of the service rate vector $\mu = (\mu_1, \dots, \mu_J)$. Then

$$\mathbf{X} \geq_{\mathcal{L}(\rho)-cc} \mathbf{X}'. \tag{8.30}$$

Example 8.4.20 *In many applications the functions in $\mathcal{F}$ serve as cost or reward functions connected with the network's performance. A typical cost function is as follows:*

Per customer at node j and per time unit a cost of amount α_j occurs, so $f_j(X_j(t)) = \alpha_j \cdot X_j(t)$ is the cost at node j. Incorporating a fixed constant cost a then in state $(n_1, \dots, n_J)$ the total cost per time unit is $f(n_1, \dots, n_J) = a + \sum_{i=1}^{J} \alpha_i \cdot n_i$. When we put the natural assumption that the costs increase when the service speed decreases, this situation is covered by the preceding theorem.

Our next theorem is in the class of generalized tandem networks as described in Proposition 8.4.17. Robin-Hood transforms under this graph structure are of the following form: Shift (probability) mass $\alpha > 0$ from arcs $r_{j,j+1}$ and $r_{j+1,j}$ to arcs $r_{j,j}$ and $r_{j+1,j+1}$. This has the following consequences.

Theorem 8.4.21 (General tandem) *Consider Jackson network processes $\tilde{\mathbf{X}}, \tilde{\mathbf{X}}'$ on state space $\mathbb{N}^J$ equipped with the partial sum order $\leq_*$, having the same stationary distribution $\tilde{\pi}^J$.*

Assume further that for some fixed $j \in \{1,\dots,J-1\}$ *and* $\alpha > 0$ *it holds* $\tilde{r}_{j(j+1)} > \alpha$ *and* $\tilde{r}_{(j+1)j} \geq \alpha$*, and that the routing for* $\tilde{\mathbf{X}}'$ *is obtained by the Robin-Hood transformation according to* (8.22)*, where* $a = b = j$ *and* $c = d = j+1$.

Then with $\mathcal{PS} := \mathcal{I}^*(\mathbb{R}^J, \leq_*) \cup \mathcal{D}^*(\mathbb{R}^J, \leq_*)$ *we have*

$$\tilde{\mathbf{X}} \leq_{\mathcal{PS}-cc} \tilde{\mathbf{X}}'. \tag{8.31}$$

It is worth mentioning that a Robin-Hood transformation applied to the tandem routing yields the Peskun ordering between the routing matrices (see Definition 8.4.4) but we do not need reversibility in the above theorem, it is substituted by the time symmetric monotonicity.

8.5 Jackson networks with breakdowns

The class of Jackson networks can be reasonably extended. Assume the servers at the nodes in a Jackson network to be unreliable, i.e., the nodes may break down. The breakdown event may occur in different ways. Nodes may break down as an isolated event or in groups simultaneously, and the repair of the nodes may end for each node individually or in groups as well. It is not required that those nodes which stopped service simultaneously return to service at the same time instant. To describe the system's evolution we have to enlarge the state space for the network process as will be described below. For a more detailed description see Sauer and Daduna [75].

8.5.1 Product formula

Control of breakdowns and repairs is as follows:

Let $\mathrm{I} \subset \mathrm{J}$ be the set of nodes in down status and $\mathrm{H} \subset \mathrm{J} \backslash \mathrm{I}, \mathrm{H} \neq \emptyset$, be some subset of nodes in up status. Then the nodes of H break down with intensity $\alpha(\mathrm{I}, \mathrm{I} \cup \mathrm{H})$.

Nodes in down status neither accept new customers nor continue serving the old customers which will wait for the server's return. (At nodes i under repair the service intensities $\mu_i(n_i)$ are set to 0). Therefore, the routing matrix has to be changed so that customers attending to join a node in down status are rerouted to nodes in up status or to the outside. We describe three possible rerouting schemes below.

Assume the nodes in I are under repair, $\mathrm{I} \subset \mathrm{J}, \mathrm{I} \neq \emptyset$. Then if $\mathrm{H} \subset \mathrm{I}, \mathrm{H} \neq \emptyset$, the nodes of H return from repair as a batch group with intensity $\beta(\mathrm{I}, \mathrm{I} \setminus \mathrm{H})$ and immediately resume their services. Routing then has to be updated again as will be described below.

The intensities for occurrence of breakdowns and repairs have to be set under constraints. A rather general versatile class is defined as follows.

Definition 8.5.1 *Let* I *be the set of nodes in down status. The intensities for breakdowns, resp. repairs for* $\mathrm{H} \neq \emptyset$ *are defined by*

$$\alpha(\mathrm{I},\mathrm{I}\cup\mathrm{H}) := \frac{a(\mathrm{I}\cup\mathrm{H})}{a(\mathrm{I})}, \quad \textit{resp.} \quad \beta(\mathrm{I},\mathrm{I}\setminus\mathrm{H}) := \frac{b(\mathrm{I})}{b(\mathrm{I}\setminus\mathrm{H})}, \tag{8.32}$$

where a and b are any functions, $a,b : \mathcal{P}(\mathrm{J}) \to [0,\infty)$ *whereas* $\frac{0}{0} := 0$.
The above intensities are assumed henceforth to be finite.

The rerouting matrices of interest are as follows.

Definition 8.5.2 (BLOCKING) *Assume that the routing matrix of the original process is reversible. Assume the nodes in* I *are the nodes of the Jackson network presently under repair. Then the routing probabilities are redefined on* $\mathrm{J}_0\setminus\mathrm{I}$ *according to*

$$\tilde{r}^{\mathrm{I}}_{ij} = \begin{cases} \tilde{r}_{ij}, & i,j \in \mathrm{J}_0\setminus\mathrm{I}, i \neq j, \\ \tilde{r}_{ii} + \sum_{k\in\mathrm{I}} \tilde{r}_{ik}, & i \in \mathrm{J}_0\setminus\mathrm{I}, i = j. \end{cases} \tag{8.33}$$

Note that even in case of $\tilde{r}_{00} = 0$, *external arrivals may be now rejected with positive probability to an immediate departure, because arrivals to nodes under repair are rerouted:*

$$\tilde{r}^{\mathrm{I}}_{00} = \tilde{r}_{00} + \sum_{k\in\mathrm{I}} \tilde{r}_{0k} \geq 0.$$

Definition 8.5.3 (STALLING) *If there is any breakdown of either a single node or a group of nodes, then all arrival streams to the network and all service processes at the nodes in up status are completely interrupted and resumed only when all nodes are repaired again.*

Definition 8.5.4 (SKIPPING) *Assume that the nodes in* I *are the nodes presently under repair. Then the routing matrix is redefined on* $\mathrm{J}_0 \setminus \mathrm{I}$ *according to:*

$$\tilde{r}^{\mathrm{I}}_{jk} = \tilde{r}_{jk} + \textstyle\sum_{i\in\mathrm{I}} \tilde{r}_{ji}\tilde{r}^{\mathrm{I}}_{ik}, \quad k,j \in \mathrm{J}_0\setminus\mathrm{I},$$

$$\tilde{r}^{\mathrm{I}}_{ik} = \tilde{r}_{ik} + \textstyle\sum_{l\in\mathrm{I}} \tilde{r}_{il}\tilde{r}^{\mathrm{I}}_{lk}, \quad i \in \mathrm{I}, k \in \mathrm{J}_0\setminus\mathrm{I}.$$

For describing the breakdown of nodes in Jackson networks we have to attach to the state spaces $\mathbb{E} = \mathbb{N}^J$ of the corresponding network processes an additional component which carries information of the reliability behavior of the system described by a process $\mathbf{Y}$. We introduce states of the form

$$(\mathrm{I}; n_1, n_2, \ldots, n_J) \in \mathcal{P}(\mathrm{J}) \times \mathbb{N}^J.$$

The meaning of such a prototype state is:

I is the set of nodes under repair. For $j \in \mathrm{J}\setminus\mathrm{I}$, the numbers $n_j \in \mathbb{N}$ indicate that at nodes j which work in a normal up status, there are n_j customers present; for $i \in \mathrm{I}$ the numbers $n_i \in \mathbb{N}$ indicate that at each node i which is in down status there are n_i

customers that wait at node i for the return of the repaired server. Collecting these states we define for the networks new Markov processes $\tilde{\mathbf{Z}} = (\mathbf{Y}, \tilde{\mathbf{X}})$ on

$$\widetilde{\mathbb{E}} = \mathcal{P}(\mathrm{J}) \times \mathbb{N}^J. \tag{8.34}$$

For such general models with breakdowns and repairs and with the above rerouting principles it was shown in Sauer and Daduna [75] that on the state space $\widetilde{\mathbb{E}}$ the steady state distribution for $\tilde{\mathbf{Z}}$ is of a product form. Note that the breakdown/repair process $\mathbf{Y}$ is Markovian on the state space $\mathcal{P}(\mathrm{J})$ of all subsets of J, but that the network process component $\tilde{\mathbf{X}}$ is in this setting **not a Markov** process.

Theorem 8.5.5 *The process $\tilde{\mathbf{Z}}$ with breakdown and repair intensities given by Eq. (8.32) and rerouting according to either* BLOCKING *or* STALLING*, or* SKIPPING *has a stationary distribution of product form given by:*

$$\tilde{\pi}^{Y,J}(\mathrm{I}; n_1, n_2, \ldots, n_J) = \pi^Y(\mathrm{I})\, \tilde{\pi}^J(n_1, n_2, \ldots, n_J)$$

with

$$\pi^Y(\mathrm{I}) = \left(1 + \sum_{\substack{\mathrm{K} \subset \mathrm{J} \\ \mathrm{K} \neq \emptyset}} \frac{a(\mathrm{K})}{b(\mathrm{K})}\right)^{-1} \frac{a(\mathrm{I})}{b(\mathrm{I})} \quad \text{for } \mathrm{I} \subset \mathrm{J}$$

and $\tilde{\pi}^J(n_1, n_2, \ldots, n_J)$ the equilibrium distribution in the standard Jackson network.

Note that time evolution of the queueing process $\tilde{\mathbf{X}}$ is different in all cases (standard Jackson, BLOCKING, STALLING, SKIPPING). At the same time, it is possible to change the breakdown/repair intensities in such a way that the stationary distribution for the joint process $\tilde{\mathbf{Z}}$ remains unchanged.

8.5.2 Bounds via dependence ordering for networks with breakdowns

8.5.2.1 Dependence ordering of Jackson networks with breakdowns

Consider Markov processes $\tilde{\mathbf{Z}} = (\mathbf{Y}, \tilde{\mathbf{X}})$ on $\widetilde{\mathbb{E}} = \mathcal{P}(\mathrm{J}) \times \mathbb{N}^J$ describing the state of the Jackson network with breakdowns. For a given set K denote by $\{\mathrm{K}\}^{\uparrow}$, $\{\mathrm{K}\}^{\downarrow}$, $\{\mathrm{K}\}^{\prec}$ the sets of its ancestors, descendants and relatives, respectively, i.e.

$$\begin{aligned} \{\mathrm{K}\}^{\uparrow} &:= \{\mathrm{I} \subseteq \mathrm{J} : \mathrm{K} \subset \mathrm{I}, \mathrm{K} \neq \mathrm{I}\}, \\ \{\mathrm{K}\}^{\downarrow} &:= \{\mathrm{I} \subseteq \mathrm{J} : \mathrm{I} \subset \mathrm{K}, \mathrm{K} \neq \mathrm{I}\}, \\ \{\mathrm{K}\}^{\prec} &:= \{\mathrm{K}\}^{\downarrow} \cup \{\mathrm{K}\}^{\uparrow}. \end{aligned}$$

Recall that $\mathbf{Y} = (Y(t), t \geq 0)$ is a cadlag Markov process on the state space $\mathcal{P}(\mathrm{J})$ which describes availability of the network's components over time, i.e. $Y(t) = \mathrm{K}$,

$\mathrm{K} \in \mathcal{P}(\mathrm{J})$, means that at time t the set K consists of the nodes which are under repair. We have

$$q^Y(\mathrm{K},\mathrm{H}) = \begin{cases} \alpha(\mathrm{K},\mathrm{H}) = \frac{a(\mathrm{H})}{a(\mathrm{K})}, & \text{if } \mathrm{H} \in \{\mathrm{K}\}_+^{\uparrow}, \\ \beta(\mathrm{K},\mathrm{H}) = \frac{b(\mathrm{K})}{b(\mathrm{H})}, & \text{if } \mathrm{H} \in \{\mathrm{K}\}^{\downarrow}, \\ -\sum_{\mathrm{I}\in\{\mathrm{K}\}^{\uparrow}} \frac{a(\mathrm{I})}{a(\mathrm{K})} - \sum_{\mathrm{I}\in\{\mathrm{K}\}^{\downarrow}} \frac{b(\mathrm{K})}{b(\mathrm{I})}, & \text{if } \mathrm{H} = \mathrm{K}, \\ 0, & \text{otherwise}. \end{cases} \tag{8.35}$$

We define for fixed $\mathrm{I}_1 \subset \mathrm{I}_2$, $\mathrm{J}_1 \subset \mathrm{J}_2$ new intensities by

$$q^{Y^\varepsilon}(\mathrm{K},\mathrm{H}) = \begin{cases} q^Y(\mathrm{K},\mathrm{H}) + \frac{\varepsilon}{\pi^Y(\mathrm{K})}, & \text{if } (\mathrm{K}=\mathrm{I}_1, \mathrm{H}=\mathrm{J}_1) \text{ or } (\mathrm{K}=\mathrm{I}_2, \mathrm{H}=\mathrm{J}_2), \\ q^Y(\mathrm{K},\mathrm{H}) - \frac{\varepsilon}{\pi^Y(\mathrm{K})}, & \text{if } (\mathrm{K}=\mathrm{I}_1, \mathrm{H}=\mathrm{J}_2) \text{ or } (\mathrm{K}=\mathrm{I}_2, \mathrm{H}=\mathrm{J}_1), \\ q^Y(\mathrm{K},\mathrm{H}), & \text{otherwise}. \end{cases} \tag{8.36}$$

Consider the processes $\mathbf{Y}$, $\mathbf{Y}^\varepsilon$ on state space $(\mathcal{P}(\mathrm{J}),\subseteq)$ and two processes $\tilde{\mathbf{Z}} = (\mathbf{Y},\tilde{\mathbf{X}})$, $\tilde{\mathbf{Z}}^\varepsilon = (\mathbf{Y}^\varepsilon,\tilde{\mathbf{X}}^\varepsilon)$ which have the same routing matrices and service intensities but different breakdown/repair processes $\mathbf{Y}$ and $\mathbf{Y}^\varepsilon$.

The following property is taken from Daduna et al [22]. Note that both processes, before and after modification, have the same product form invariant distribution, but they are different in their time evolution. The modification results in a higher rate to change sets under repair to "similar" ones. Of course such a transformation can be iterated, which leads to eliminate transitions between not ordered sets. Note that processes under comparison are not Markovian (the "big" process $\mathbf{Z}$ is Markovian, but $\tilde{\mathbf{X}}$ usually not).

Property 8.5.6 (Enlarging dependence in time via structure of breakdowns) *Assume that two Jackson networks have the same arrival intensities, the same rerouting matrices according to either* BLOCKING *or* STALLING *or* SKIPPING *and breakdown/repair intensity matrices are given by* (8.35) *and* (8.36)*. Assume also that breakdown/repair intensity matrices and its time-reversal counterparts are stochastically monotone. Then in equilibrium, for all $n \geq 2$ and $t_1 \leq \cdots \leq t_n$,*

$$E\left[f\left(\tilde{X}(t_1),\ldots,\tilde{X}(t_n)\right)\right] \leq E\left[f\left(\tilde{X}^\varepsilon(t_1),\ldots,\tilde{X}^\varepsilon(t_n)\right)\right],$$

for all functions f with isotone differences on $(\widetilde{\mathbb{E}}^n, (\leq^J)^n)$.

8.6 General networks

Consider an open network of J, k_j-server, FCFS nodes, $j \in \mathrm{J} = \{1,\ldots,J\}$. We set $k = (k_1,\ldots,k_J)$. Denote by $N^0 = (N^1,\ldots,N^J)$ the vector of counting processes of arrivals from outside to the nodes, by $S = (S^1,\ldots,S^J)$ the vector of service time sequences $S^j = (S^j_1,\ldots)$, where S^j_n denotes the service time received by the n-th initiated job at station j. Denote by $V = (V^1,\ldots,V^J)$ the vector of destination sequences $V^j = (V_1^j, \ldots)$, where V_n^j denotes the number of the node visited by the job that is the n-th departing from the node j or $V_n^j = 0$ if the job leaves the network. Let $\tilde{\mathbf{X}} = (\tilde{X}(t) : t \geq 0)$ denote the vector process recording the joint queue lengths in the network for time t. For $t \in \mathbb{R}_+$, $\tilde{X}(t) = (\tilde{X}_1(t), \ldots, \tilde{X}_J(t))$ means that at time t there are $\tilde{X}_j(t)$ customers present at node j, either in service or waiting. Given an initial content $\tilde{X}(0) = (\tilde{X}_1(0), \ldots, \tilde{X}_J(0))$, such a general network is determined by the arrival, service and routing variables and will be denoted therefore by $(N^0, V)/S, k/J$. The corresponding closed network, which starts with N customers and does not admit arrivals from outside will be denoted by $V/S, k/J + N$. Denote by $N^d = (N^{1,\cdot}, \ldots, N^{J,\cdot})$ the vector of point processes of departures from the nodes, and by $N^a = (N^{\cdot,1}, \ldots, N^{\cdot,J})$ the vector of all arrivals to the nodes. The limits (if they exist) $\lim_{t\to\infty} N^{j,\cdot}(t)/t$, which are the throughputs of the consecutive nodes will be denoted by $TH_j(V/S, k/J + N)$, $j \in \mathrm{J}$.

For an open network of J, k_j-server, FCFS nodes, with finite waiting rooms of sizes $B_1, \ldots, B_J$ we introduce additional parameter $B = (B_1, \ldots, B_j)$ and use notation $(N^0, V)/S, k, B/J$ for open networks, and $V/S, k, B/J + N$ for closed networks. An arriving job from outside that finds the selected node full is lost. A job that completes service in node j proceeds to the next node according to V^j unless the latter is full. In this case we consider *manufacturing blocking*: the job has to wait until there is an empty space in the selected node, i.e. the server at node j is idle (blocked); or we consider *communication blocking*: if a job completes service at j and finds the next node full, it has to repeat service at j.

An alternative description of a J-variate arrival process is the one given by a sequence

$$\Phi \equiv \{(T_n^1, \ldots, T_n^J)\}_{n=-\infty}^{\infty}$$

of random variables defined on a probability space $(\Omega, \mathcal{F}, P)$, such that $T_0^i \leq 0 < T_1^i$, $T_n^i < T_{n+1}^i$, $i = 1, \ldots, J$, $n \in \mathbb{Z}$ and $\lim_{n\to\pm\infty} T_n^i = \pm\infty$ (Φ is nonexplosive). Denote by $\{X_n^i\}_{n=-\infty}^{\infty}$ a sequence of inter-point distances, i.e. $X_n^i = T_n^i - T_{n-1}^i$ (the interval X_1^i contains 0). Then a J-variate point process Φ can be seen as a random element assuming its values in $(\mathbb{R}_+^\infty)^J$.

Let $\mathcal{N}$ be a set of locally finite integer valued measures on $\mathbb{R}$. Equivalently, we view Φ as a random measure $\Phi : \Omega \to \mathcal{N}^k$ with the coordinate functions $\Phi = (\Phi^1, \ldots, \Phi^k)$, $\Phi^i : \Omega \to \mathcal{N}$. Then for all Borel sets B, $N_\Phi^i(B) := \Phi^i(B)$ is the corresponding counting variable. However, if it is clear which point process do we mean we shall write shortly N^i instead of N_Φ^i. The corresponding counting processes $(N^i(t), t \geq 0)$, $i = 1, \ldots, J$ are given by $N^i(t) := N^i((0,t])$.

It will be convenient to have notation for another point process Ψ with the corresponding points $\{(\mathcal{T}_n^1, \ldots, \mathcal{T}_n^k)\}_{n\geq 1}$, $k \leq \infty$ and inter-point distances $U_n^i = \mathcal{T}_n^i - \mathcal{T}_{n-1}^i$, $i = 1, \ldots, k$.

In the case $k = 1$ we shall write T_n (X_n, N, λ) and $\mathcal{T}_n$ (U_n) instead of writing these quantities with the superscript 1.

We denote by $\mathcal{L}_{\mathrm{st}}$ ($\mathcal{L}_{\mathrm{cx}}$, $\mathcal{L}_{\mathrm{icx}}$) the class of increasing (convex, increasing and convex) functions $f : \mathbb{R} \to \mathbb{R}$.

Define for $1 \leq l \leq m$, $\varepsilon > 0$ and arbitrary function $\varphi : \mathbb{R}^m \to \mathbb{R}$ the difference operator Δ_l^ε by

$$\Delta_l^\varepsilon \varphi(u_1,\ldots,u_m) = \varphi(u_1,\ldots,u_{l-1},u_l+\varepsilon,u_{l+1},\ldots,u_m) - \varphi(u_1,\ldots,u_m)$$

for given $u_1,\ldots,u_m$.

We denote arbitrary m-dimensional intervals by $\mathfrak{J} \subseteq \mathbb{R}^m$, i.e. $\mathfrak{J} = I^1 \times \cdots \times I^m$, where I^j is a (possibly infinite ended) interval on $\mathbb{R}$ for $j = 1,\ldots,m$. Recall that a function $\varphi : \mathbb{R}^m \to \mathbb{R}$ is *supermodular* on $\mathfrak{J}$ if for all $1 \leq l < j \leq m$, $\varepsilon_l, \varepsilon_j > 0$ and $\mathbf{u} = (u_1,\ldots,u_m) \in \mathfrak{J}$ such that $(u_1,\ldots,u_{l-1},u_l+\varepsilon_l,u_{l+1},\ldots,u_m) \in \mathfrak{J}$ we have

$$\Delta_l^{\varepsilon_l}\Delta_j^{\varepsilon_j}\varphi(\mathbf{u}) \geq 0\,.$$

A function $\varphi : \mathbb{R}^m \to \mathbb{R}$ is *directionally convex* on $\mathfrak{J}$ if it is supermodular on $\mathfrak{J}$ and convex w.r.t. each coordinate on I^j, $j = 1,\ldots,m$ or, equivalently

$$\Delta_l^{\varepsilon_l}\Delta_j^{\varepsilon_j}\varphi(\mathbf{u}) \geq 0$$

for all $1 \leq l \leq j \leq m$. We denote by $\mathcal{L}_{\mathrm{sm}}(\mathfrak{J})$ ($\mathcal{L}_{\mathrm{dcx}}(\mathfrak{J})$) the class of all supermodular (directionally convex) functions on $\mathfrak{J}$. Moreover, we denote the class of increasing directionally convex functions on $\mathfrak{J}$ by $\mathcal{L}_{\mathrm{idcx}}(\mathfrak{J})$ and symmetric supermodular functions on $\mathfrak{J}$ by $\mathcal{L}_{\mathrm{ssm}}(\mathfrak{J})$. We skip $\mathfrak{J}$ in this notation if $\mathfrak{J} = \mathbb{R}^m$.

For arbitrary random vectors $(Y_1,\ldots,Y_n)$, $(\tilde{Y}_1,\ldots,\tilde{Y}_n)$ defined on probability spaces $(\Omega,\mathfrak{F},P)$ and $(\tilde{\Omega},\tilde{\mathfrak{F}},\tilde{P})$ respectively, we write

$$(Y_1,\ldots,Y_n) <_{\mathrm{a}} (\tilde{Y}_1,\ldots,\tilde{Y}_n)$$

if

$$E[\varphi(Y_1,\ldots,Y_n)] \leq E[\varphi(\tilde{Y}_1,\ldots,\tilde{Y}_n)],$$

for all $\varphi : \mathbb{R}^n \to \mathbb{R}$ such that $\varphi \in \mathcal{L}_{\mathrm{a}}$, where $\mathcal{L}_{\mathrm{a}}$ denotes one of the classes $\mathcal{L}_{\mathrm{sm}}$, $\mathcal{L}_{\mathrm{dcx}}$, $\mathcal{L}_{\mathrm{idcx}}$. Similarly, for random sequences $\{Y_n\}_{n\geq 1}$ and $\{\tilde{Y}_n\}_{n\geq 1}$ we write $\{Y_n\} <_{\mathrm{a}} \{\tilde{Y}_n\}$ if for all $n \geq 1$, $(Y_1,\ldots,Y_n) <_{\mathrm{a}} (\tilde{Y}_1,\ldots,\tilde{Y}_n)$.

Let Ψ ($\tilde{\Psi}$) be a J-variate stationary point process with the corresponding inter-point distances $\{U_n^i\}$ ($\{\tilde{U}_n^i\}$), $i = 1,\ldots,k$. We write

- $\Psi <_{\mathrm{m-a-\infty}} \tilde{\Psi}$ if $(\{U_n^1\},\ldots,\{U_n^J\}) <_{\mathrm{a}} (\{\tilde{U}_n^1\},\ldots,\{\tilde{U}_n^J\})$, i.e. if for all $n \geq 1$, ,

$$\left((U_1^1,\ldots,U_n^1),\ldots,(U_1^J,\ldots,U_n^J)\right) <_{\mathrm{a}} \left((\tilde{U}_1^1,\ldots,\tilde{U}_n^1),\ldots,(\tilde{U}_1^J,\ldots,\tilde{U}_n^J)\right)\,.$$

Let Φ ($\tilde{\Phi}$) be a J-variate point process with the corresponding counting measures N^i ($\tilde{N}^i$), $i = 1,\ldots,J$. We write

- $\Phi <_{\mathrm{m-a-D}} \tilde{\Phi}$ if for all $0 \leq t_1 < t_2 < \cdots < t_r$, $r \geq 1$,

$$(N^i(t_1),\ldots,N^i(t_r), i = 1,\ldots,J) <_{\mathrm{a}} (\tilde{N}^i(t_1),\ldots,\tilde{N}^i(t_r), i = 1,\ldots,J)\,.$$

Let $\mathfrak{I} = \{I_n\}_{n\geq 1}$ be a partition of $\mathbb{R}_+$ such that I_r, $r \geq 1$ have the same length. We write

- $\Phi <_{\mathrm{m-a-}\mathcal{N}} \tilde{\Phi}$ if for all $(I_1,\ldots,I_r)$, $r \geq 1$,

$$(N^i(I_1),\ldots,N^i(I_r), i=1,\ldots,J) <_{\mathrm{a}} (\tilde{N}^i(I_1),\ldots,\tilde{N}^i(I_r), i=1,\ldots,J).$$

Here $<_{\cdot\cdot\infty}$ ($<_{\cdot\cdot\mathcal{N}}$, $<_{\cdot\cdot\mathrm{D}}$) stands for the comparison of point processes considered as random elements of $(\mathbb{R}_+^{\infty})^J$, $(\mathcal{N}^k$, $(\mathrm{D}([0,\infty)))^k J)$, where $\mathrm{D}([0,\infty))$ is the space of right-hand-side continuous functions with left-hand-side limits.
For 1-variate point processes ($J=1$) we shall omit subscript 1, and write $<_{\mathrm{a-D}}$, $<_{\mathrm{a-}\mathcal{N}}$, $<_{\mathrm{a-}\infty}$ coincides with orderings defined in Kwieciński and Szekli [48].

8.6.1 Dependence and variability in input

The next property proved by Meester and Shanthikumar [62] is a general result connected with so called Ross's conjecture, which still receives some attention in the context of single queues.

Property 8.6.1 *Consider two open networks with finite waiting rooms $(N^0,V)/S,1,B/J$, and $(N'^0,V)/S,1,B/J$ which operate according to the manufacturing blocking (1 denotes the vector with 1 on each coordinate). Assume that in N^0, and N'^0 only the first coordinates are non-trivial, and $V^j = (j+1, j+1,\ldots)$, i.e. these networks are open tandems. If S is a vector of independent sequences of independent exponential random variables with rates $\mu_j(k)$ when there are k jobs at station j which are increasing and concave functions in k then $N^{0,1} <_{idcx-\mathcal{N}} N'^{0,1}$ implies that*

$$(N^l(t), N^l(t)+\tilde{X}_1(t),\ldots,N^l(t)+\tilde{X}_1(t)+\cdots+\tilde{X}_J(t)) <_{idcx}$$

$$(N'^l(t), N'^l(t)+\tilde{X}'_1(t),\ldots,N'^l(t)+\tilde{X}'_1(t)+\cdots+\tilde{X}'_J(t),$$

where N^l denotes the point process of lost jobs.

Chang et al. [13] considered a special case where the authors assumed infinite buffers, and doubly stochastic Poisson input point process N^1, obtaining this result only for the number of jobs. Moreover for finite buffers they obtained the result for the number of lost jobs. For a more recent research of this type, where the arrival stream consists of multiple on-off sources, see Koole and Liu [43].

8.6.2 Comparison of workloads

Assume that for the routing vector $V = (V^1,\ldots,V^J)$, we have $V_n^j = 0$ for all $j \in$ J, and $n \in N$. That is the arrivals are routed to one of the J queues with infinite waiting room and after receiving service depart from the system. Arrivals are characterized by $N^0 = (N^1,\ldots,N^J)$ which can be seen as a marked point process (τ_n, Z_n), where

τ_n denotes the epoch of the nth arrival and Z_n denotes the number of the station this arrival is routed to. Consider a parallel system with *resequencing synchronization*, which means that the nth customer departs from the system provided that all the customers that arrived earlier have been served. Denote by $W(t) = (W^1(t),\dots,W^J(t))$ the amount of work in the queues at time t. The next property comes from Chang Cheng-Shang [12].

Property 8.6.2 *Suppose that in a parallel system described above, (Z_n) is a stationary Markov chain independent of (τ_n) and S, with the transition probabilities $P(Z_{n+1} = j \mid Z_n = i) = (1-\sigma)/J$, $i \neq j$, and $P(Z_{n+1} = i \mid Z_n = i) = \sigma + (1-\sigma)/J$, for some parameter $\sigma \in [0,1]$. Then for each t, $E(f(W(t)))$ is increasing as a function of σ, provided f is coordinate-wise increasing, symmetric, submodular, and convex in each variable.*

8.6.2.1 Workload in parallel queues

Consider a queueing system of J parallel $G/G/1$ FIFO queues. The input is generated by kJ-variate point processes Φ (interarrival times) and Ψ (service times), independent of Φ. For $t \geq 0$ and $I = (a,b]$ define

$$M^i(t) = \sum_{n=1}^{N^i(t)} U^i_n, \quad i = 1,\dots,J$$

and

$$M^i(I) = \sum_{n=N^i(a)+1}^{N^i(b)} U^i_n, \quad i = 1,\dots,J.$$

Call M^i, $i = 1,\dots,k$ cumulative processes. Denote by

$$\mathbf{W}(t) \equiv (W^1(t),\dots,W^J(t))$$

the vector of transient workloads, which is known to fulfill

$$W^i(t) = \max_{0 \leq u \leq t}(0, M^i(t) - M^i(u) - (t-u))$$

(Borovkov [6, p. 23]). Similarly, for J-variate point processes Φ', Ψ' define

$$M'^i(t) = \sum_{n=1}^{N'^i(t)} U'^i_n, \quad i = 1,\dots,J$$

and as above $M'^i(I)$ and $\mathbf{W}'(t)$. The following property is taken from Kulik and Szekli [47].

Property 8.6.3 (i) *Assume that $\boldsymbol{\Phi} <_{\text{m-idcx}-\mathcal{N}} \boldsymbol{\Phi}'$, $\boldsymbol{\Psi} = \boldsymbol{\Psi}'$ and $\boldsymbol{\Psi}$ consists of mutually independent iid sequences. Then for all* $0 < t_1 < \cdots < t_r$,

$$(\mathbf{W}(t_1), \ldots, \mathbf{W}(t_r)) <_{\text{idcx}} (\mathbf{W}'(t_1), \ldots, \mathbf{W}'(t_r)).$$

(ii) *Assume that $\boldsymbol{\Psi} <_{\text{m-idcx}-\infty} \boldsymbol{\Psi}'$, $\boldsymbol{\Phi} = \boldsymbol{\Phi}'$. Then for all* $0 < t_1 < \cdots < t_r$,

$$(\mathbf{W}(t_1), \ldots, \mathbf{W}(t_r)) <_{\text{idcx}} (\mathbf{W}'(t_1), \ldots, \mathbf{W}'(t_r)).$$

8.6.2.2 Workload in batch queues

Consider a queueing system of J parallel $G/GI/1$ FIFO queues. The input is generated by J-variate point processes $\boldsymbol{\Phi}$ (arrival times) and $\boldsymbol{\Psi}$ (batch sizes), independent of $\boldsymbol{\Phi}$. For $t \geq 0$ and $I = (a, b]$ define

$$K^i(t) = \sum_{n=1}^{N^i(t)} U_n^i, \quad i = 1, \ldots, J,$$

and

$$K^i(I) = \sum_{n=N^i(a)+1}^{N^i(b)} U_n^i, \quad i = 1, \ldots, kJ.$$

Here, $K^i(t)$ represents the number of jobs brought to a queue i up to time t. For $\{S_n^i\}_{n\geq 1}$, $i = 1, \ldots, J$, iid mutually independent service times, independent of $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ define cumulative processes

$$M^i(t) = \sum_{n=1}^{K^i(t)} S_n^i, \quad i = 1, \ldots, J,$$

and

$$M^i(I) = \sum_{n=K^i(a)+1}^{K^i(b)} S_n^i, \quad i = 1, \ldots, J.$$

Then the transient workload is given by

$$W^i(t) = \max_{0 \leq u \leq t} (0, M^i(t) - M^i(u) - (t - u)).$$

Denote by

$$\mathbf{W}(t) \equiv (W^1(t), \ldots, W^J(t))$$

the vector of transient workload. Similarly, having arrival process $\boldsymbol{\Phi}' = \boldsymbol{\Phi}$, batch size process $\boldsymbol{\Psi}'$ and the same service times, we define $K'^i(t)$, $K'^i(I)$, $M'^i(t)$, $M'^i(I)$, $W'^i(t)$ and $\mathbf{W}'(t)$.

From Kulik and Szekli [47] we have

Property 8.6.4 *Assume that* $\{(U_n^1,\dots,U_n^J)\}_{n\geq 1}$, $\{(U'^1_n,\dots,U'^J_n)\}_{n\geq 1}$ *are sequences of independent random variables such that for all* $n \geq 1$, $(U_n^1,\dots,U_n^J) <_{\mathrm{sm}} (U'^1_n,\dots,U'^J_n)$*. Then for all* $0 < t_1 < \cdots < t_r$,

$$(\mathbf{W}(t_1),\dots,\mathbf{W}(t_r)) <_{\mathrm{idcx}} (\mathbf{W}'(t_1),\dots,\mathbf{W}'(t_r))\,.$$

The assumptions in the above properties can be described in a more detailed way. Let Φ, Φ' be J-variate arrival processes with interarrival times X_n^i, X'^i_n, $i=1,\dots,J$. If $\{X_n^1,\dots,X_n^J\}_{n\geq 1}$ and $\{X'^1_n,\dots,X'^J_n\}_{n\geq 1}$ are sequences of independent random vectors and for all $n \geq 1$,

$$(X_n^1,\dots,X_n^J) <_{\mathrm{sm}} (X'^1_n,\dots,X'^J_n),$$

then $\Phi <_{\mathrm{m-sm-}\mathcal{N}} \Phi'$ (Li and Xu [53]). Assume that $X_n =^d X_n^i =^d X_n^j$, $i,j=1,\dots,J$, $n \geq 1$. From Lorentz inequality one obtains that $(X_n^1,\dots,X_n^J) <_{\mathrm{sm}} (X_n,\dots,X_n)$. Therefore, synchronization give the upper bound (in $<_{\mathrm{sm}}$ and hence in $<_{\mathrm{idcx}}$-order) for arrival processes and hence, using previous results, for workload in parallel queues.

8.6.3 Throughput in general networks

For general networks results about throughput were obtained by Shanthikumar and Yao [84], and by Tsoucas and Walrand [94]. Since the formulations of the following properties are self-explaining we shall skip comments on them.

Property 8.6.5 *Consider two general closed networks* $V/S,k/J+N$ *with an independent initial content* $X(0)$ *and* $V/S,k/J+N'$ *with an independent initial content* $X'(0)$ *such that* $X(0) \leq_{st} X'(0)$*. Then* $N^a <_{st\text{-}\mathfrak{D}} N'^a$, $N^d <_{st\text{-}\mathfrak{D}} N'^d$*, and*

$$TH_j(V/S,k/J+N) \leq TH_j(V/S,k/J+N'),\ j \in \mathrm{J}.$$

Property 8.6.6 *Consider two general closed networks* $V/S,k/J+N$ *with an initial content* $X(0)$ *and* $V/S',k/J+N$ *with equal initial content such that service time sequences are independent of the initial content and of* V*, and* $S \geq_{st} S'$*. Then* $N^a <_{st\text{-}\mathfrak{D}} N'^a$, $N^d <_{st\text{-}\mathfrak{D}} N'^d$*, and*

$$TH_j(V/S,k/J+N) \leq TH_j(V/S',k/J+N),\ j \in \mathrm{J}.$$

Property 8.6.7 *Consider two general closed networks* $V/S,k/J+N$ *with an initial content* $X(0)$ *and* $V/S,k'/J+N$ *with equal initial content such that* $k \geq k'$*. Then* $N^a <_{st\text{-}\mathfrak{D}} N'^a$, $N^d <_{st\text{-}\mathfrak{D}} N'^d$*, and*

$$TH_j(V/S,k/J+N) \leq TH_j(V/S,k'/J+N),\ j \in \mathrm{J}.$$

From Tsoucas and Walrand [94] we have

Property 8.6.8 *Consider two open networks with finite waiting rooms* $(N^0,V)/S,k,B/J$, *and* $(N^0,V)/S,k',B'/J$ *which operate according to the manufacturing blocking. Assume that in* N^0 *only the first coordinate is non-trivial, and* $V^j = (j+1, j+1,\ldots)$, *i.e. these networks are open tandems. If* N^0 *and* S *are independent and* $k \le k'$ *and* $B \le B'$ *then*

$$N^{acc} <_{st\text{-}\mathfrak{D}} N'^{acc},$$

where N^{acc} *denotes the point process of accepted jobs to the tandem.*

From Meester and Shanthikumar [61], also Anantharam, Tsoucas [3] we have

Property 8.6.9 *Consider open network with finite waiting rooms* $(N^0,V)/S,1,B/J$, *which operates according to the manufacturing blocking (*1 *denotes the vector with 1 on each coordinate). Assume that in* N^0 *only the first coordinate is non-trivial, and* $V^j = (j+1, j+1,\ldots)$, *i.e. these network is an open tandem. If* S *is a vector of independent sequences of iid exponential random variables, and* $B_1 = \infty$ *then the throughput of this tandem is increasing and concave as a function of* B.

References

1. Adan, I. and van der Wal, J. Monotonicity of the throughput of a closed queueing network in the number of jobs. *Operations Research*, 37: 953-957, 1989.
2. Alberti, P.M. and Uhlmann, A. *Stochasticity and Partial Order*. VEB Deutscher Verlag der Wissenschaften, Berlin, 1982.
3. Anantharam, V. and Tsoucas, P. Stochastic Concavity of Throughput in Series of Queues with Finite buffers. *Advances in Applied Probability*, 22: 761–763, 1990.
4. Baccelli, F. and Bremaud, P. *Elements of queueing theory. Palm martingale calculus and stochastic recurrences*. Second edition. Applications of Mathematics 26. Springer-Verlag, Berlin, 2003.
5. Bäuerle, N. and Rolski, T. A monotonicity result for the workload in Markov-modulated queues. *Journal of Applied Probability*, 35:741–747, 1998.
6. Borovkov, A. A.. *Stochastic Processes in Queueing Theory*, Wiley, 1976.
7. Boxma, O.J. and Daduna, H. Sojourn times in queueing networks. *Stochastic Analysis of Computer and Communication Systems*, Takagi, H. ed., pp. 401–450, North–Holland, Amsterdam, 1990.
8. Boxma, O J., Kelly, F. P. and Konheim, A. G. The product form for sojourn time distributions in cyclic exponential queues. *JACM*, 31: 128–133, 1984.
9. Burke, P.J. The output of a queueing system. *Operations Research*, 4: 699–704, 1956.
10. Burke, P.J. The output process of a stationary M/M/s queueing system. *Annals of Mathematical Statistics*, 39: 1144–1152, 1968.
11. Burke, P.J. The dependence of sojourn times in tandem M/M/s queues. *Operations Research*, 17: 754–755, 1969.
12. Chang, Cheng-Shang. A new ordering for stochastic majorization: theory and applications. *Advances in Applied Probability*, 24: 604-634,1992.
13. Chang, Cheng-Shang, Chao, Xiuli amd Pinedo, M. Monotonicity results for queues with doubly stochastic Poisson arrivals: Rosss conjecture. *Advances in Applied Probability*, 23: 210-228, 1991.

14. Chen, Hong and Yao, D. D. *Fundamentals of Queueing Networks. Performance, Asymptotics, and Optimization*. Springer-Verlag, New York, 2001.
15. Chen, Mu-Fa. *From Markov Chains to Non-equilibrium Particle Systems.* World Scietific, Singapore, 2004.
16. Chow, Wee - Min. The cycle time distribution of exponential cyclic queues. *JACM*, 27: 281–286, 1980.
17. Christofides, T. C. and Vaggelatou, E. A connection between supermodular ordering and positive/negative association. *Journal of Multivariate Analysis*, 88: 138–151, 2004.
18. Daduna, H., and Szekli, R. Dependencies in Markovian networks. *Advances in Applied Probability*, 27: 226–254, 1995.
19. Daduna, H. and Szekli, R. On the correlation of sojourn times in open networks of exponential multiserver queues. *Queueing Systems Theory Appl.*, 34: 169-181, 2000.
20. Daduna, H. and Szekli, R. Dependence structure of sojourn times via partition separated ordering. *Operations Research Letters*, 31: 462-472, 2003.
21. Daduna, H. and Szekli, R. On the correlation structure of closed queueing networks. *Stochastic Models*, 20: 1-29, 2004.
22. Daduna, H., Kulik, R., Sauer, C. and Szekli, R. Dependence ordering for queuing networks with breakdown and repair. *Probability in the Engineering and Informational Sciences*, 20: 575-594, 2006.
23. Daduna, H. and Szekli, R. Dependencies in Markovian networks. *Advances in Applied Probability*, 27:226–254, 1995.
24. Daduna, H. and Szekli, R. Dependence ordering for Markov processes on partially ordered spaces. *Journal of Applied Probability*, 43:793–814, 2006.
25. Daduna, H. and Szekli, R. Impact of routeing on correlation strength in stationary queueing network processes, *Journal of Applied Probability*, 45: 846-878, 2008.
26. Economou, A. Necessary and sufficient conditions for the stochastic comparison of Jackson networks. *Probability in the Engineering and Informational Sciences*, 17: 143–151, 2003.
27. Economou, A. On the stochastic domination for batch-arrival, batch-service and assemble-transfer queueing networks. *Journal of Applied Probability*, 40: 1103-1120, 2003.
28. Foley, R. D. and Kiessler, P. C. Positve correlations in a three–node Jackson queueing network. *Advances of Applied Probability*, 21: 241–242, 1989.
29. Fry, T.C. Probability and Its Engineering Uses. Princeton, N. Y.: Van Nostrand, 1928.
30. Glasserman, P. and Yao, D. D. *Monotone structure in discrete-event systems*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley and Sons, New York, 1994.
31. Gordon, W.J. and Newell, G.F. Closed queueing networks with exponential servers. *Operations Research*, 15:254–265, 1967.
32. Gross, D. and Harris, C.M. Fundamentals of Queueing Theory. New York: John Wiley and Sons, 1974.
33. Harris, T. G. A correlation inequality for Markov processes in partially ordered spaces. *Annals of Probability*, 5:451–454, 1977.
34. Hu, T. and Pan, X. Comparisons of dependence for stationary Markov processes. *Probability in the Engineering and Informational Sciences*, 14: 299–315, 2000.
35. Hu, T. , Müller, A. and Scarsini, M. Some counterexamples in positive dependence. *Journal of Statistical Planning and Inference*, 124: 153 – 158, 2004.
36. Jackson, J. R. Networks of waiting lines. *Operations Research*, 4: 518–521, 1957.
37. Joe, H. *Mulivariate Models and Dependence Concepts*. Chapman and Hall, London, 1997.
38. Kanter, M. Lower bounds for the probability of overload in certain queueing networks. *Journal of Applied Probability*, 22: 429–436, 1985.
39. Keilson, J. and Kester, A. Monotone matrices and monotone Markov processes. *Stochastic Processes and Their Applications*, 5: 231–241, 1977.

40. Kelly, F. P. Reversibility and Stochastic Networks. New York: Wiley, 1979.
41. Kelly, F. and Pollett, P. Sojourn times in closed queueing networks. *Advances in Applied Probability*, 15: 638–653, 1983.
42. Khintchine A. Y. Mathematical theory of a stationary queue. *Mat. Sbornik* **39**, 73-84, 1932.
43. Ger Koole, Zhen Liu. Stochastic Bounds for Queueing Systems with Multiple On-Off Sources. *Probability in the Engineering and Informational Sciences*, 12: 25–48, 1998.
44. Heyman, D. P. and Sobel, M. J. *Stochastic models in operations research*, Vol. II.
45. Iscoe, I. and McDonald, D.I. Asymptotics of exit times for Markov jump processes I. *Annals of Applied Probability*, 22:372–397, 1994.
46. Iscoe, I. and McDonald, D. Asymptotics of exit times for Markov jump processes II: Applications to Jackson networks. *Annals of Applied Probability*, 22:2168–2182, 1994.
47. Kulik, R. and Szekli, R. Dependence orderings for some functionals of multivariate point processes. *Journal of Multivariate Analysis*, 92: 145–173, 2005.
48. Kwieciński A. and Szekli, R. A compensator conditions for stochastic ordering of point processes, *J. Appl. Prob.*, 28: 751–761, 1991.
49. Last, G. and Brandt, A. *Marked Point Processes on the Real Line. The Dynamic Approach*. Springer-Verlag, New York, 1995.
50. Liggett, T. M.. *Interacting Particle Systems* Springer-Varlag, New York, 1985.
51. Lindvall, T. Stochastic monotonicities in Jackson queueing networks. *Probability in the Engineering and Informational Sciences*, 11: 1–9, 1997.
52. Kelbert M., Kontsevich M.L., Rybko A.N. On Jacksons networks on denumerable graphs. *Theory of Probability and Applications*, 33: 379–382,1988.
53. Li, H. and Xu, S.H. On the dependence structure and bounds of correlated parallel queues and their applications to synchronized stochastic systems, *Journal of Applied Probability*, 37: 1020–1043, 2000.
54. Li, H. and Shaked, M. . Stochastic convexity and concavity of Markov processes. *Mathematics of Operations Research*, 19:477–493, 1994.
55. Li, H. and Xu, S. H. Stochastic bounds and dependence properties of survival times in a multicomponent shock model. *Journal of Applied Probability*, 37:1020–1043, 2000.
56. Lorek, P. *Speed of convergence to stationarity for stochastically monotone Markov chains*. PhD thesis, Mathematical Institute, University of Wroclaw, 2007.
57. Marshall, A. W. and Olkin, I. *Inequalities: Theory of Majorisation and Its Applications*. Academic Press, New York, 1979.
58. Massey, W. A. An Operator-Analytic Approach to the Jackson Network. *Journal of Applied Probability*, 21: 379–393, 1984.
59. Massey, W. A.. A Family of Bounds for the Transient Behavior of a Jackson Network. *Journal of Applied Probability*, 23: 543–549, 1986.
60. Massey, W. A.. Stochastic ordering for Markov processes on partially ordered spaces. *Mathematics of Operations Research*, 12: 350–367, 1987.
61. Meester, L. E. and Shanthikumar, J. G. Concavity of the throughput of tandem queueing systems with finite buffer storage space. *Advances in Applied Probability*, 22: 764-767, 1990.
62. Meester, L. E. and Shanthikumar, J. G. Regularity of stochastic processes: a theory based on directional convexity. *Probability in the Engineering and Informational Sciences*, 7: 343–360, 1993.
63. Melamed, B. Sojourn times in queueing networks. *Mathematics of Operations Research*, 7: 223–244, 1982.
64. Molina, E. C. Application of the Theory of Probability to Telephone Trunking Problems. *Bell Syst. Tech. J.***6**, 461-494, 1927.
65. Müller, A. and Stoyan, D. *Comparison Methods for Stochastic Models and Risks*. Wiley, Chichester, 2002.
66. Newell, G. F. Applications of Queueing Theory. London: Chapman and Hall Ltd., 1971.
67. Palm, C. Analysis of the Erlang Traffic Formulae for Busy-Signal Arrangements. *Ericsson Tech.***6**, 39-58, 1938.

68. Peskun, P.H. Optimum Monte Carlo sampling using Markov chains. *Biometrika*, 60:607-612, 1973.
69. Pestien, V. and Ramakrishnan, S. Monotonicity and asymptotic queue-length distribution in discrete-time networks. *Queueing Systems*, 40: 313-331, 2002.
70. Pollaczek F. Über eine Aufgabe der Wahrscheinlichkeitstheorie. *Math. Zeit.* **32** 64-100 and 729-750, 1930.
71. Reich, E. Waiting times when queues are in tandem. *Annals of Mathematical Statistics*, 28: 768–773, 1957.
72. Reich, E. Note on queues in tandem. *Annals of Mathematical Statistics*, 34: 338–341, 1963.
73. Ridder,A.A. *Stochastic Inequalities for Queues*. Thesis, Leiden.
74. Rüschendorf, L. Comparison of multivariate risks and positive dependence. *Advances in Applied Probability*, 41: 391–406, 2004.
75. Sauer C. and Daduna, H. Availability Formulas and Performance Measures for Separable Degradable Networks. *Economic Quality Control*, 18: 165–194.
76. Serfozo, R. Introduction to Stochastic Networks. Springer, 1999.
77. Shaked, M. and Shanthikumar, J. G. *Stochastic Orders*. Springer Series in Statistics. Springer, New York, 2007.
78. Shanthikumar, J. G. Stochastic majorization of random variables with proportional equilibrium rates. *Advances in Applied Probability*, 19: 854-872, 1987.
79. Shanthikumar, J.G. and Yao, D.D. The effect of increasing service rates in closed queueing network. *Journal of Applied Probability*, 23: 474 – 483, 1986.
80. Shanthikumar, J. G. and Yao, D.D. Stochastic monotonicity of the queue lengths in closed queueing networks. *Operations Research*, 35: 583-588, 1987.
81. Shanthikumar, J.G. and Yao, D.D. Throughput bounds for closed queueing networks with queue-dependent service rates. *Performance Evaluation* , 9: 69-78, 1988/89.
82. Shanthikumar, J.G. and Yao, D.D. Second order properties of the throughput of a closed queueing network. *Mathematics of Operations Research*, 13: 524-534, 1988.
83. Shanthikumar, J.G. and Yao, D.D. Monotonicity and concavity properties in cyclic queueing networks with finite buffers. *Queueing networks with blocking.* (Raleigh, NC, 1988), 325-344, North-Holland, Amsterdam, 1989.
84. Shanthikumar, J. G. and Yao, D.D. Stochastic monotonicity in general queueing networks. *Journal of Applied Probability*, 26: 413-417, 1989.
85. Schassberger, R. and Daduna H. The time for a roundtrip in a cycle of exponential queues. *JACM*, 30: 146–150, 1983.
86. Schassberger, R. and Daduna, H. Sojourn times in queueing networks with multiserver nodes. *Journal of Applied Probability*, 24: 511–521, 1987.
87. Simon, B. and Foley, R. D. Some results on sojourn times in acyclic Jackson networks. *Management Sciences*, 25: 1027–1034, 1979.
88. Szekli, R. *Stochastic Ordering and Dependence in Applied Probability*. Lecture Notes in Statistics 97, Springer-Verlag, 1995.
89. Tierney, L. A note on Metropolois-Hastings kernels for general state spaces. *Annals of Applied Probability*, 8:1-9, 1998.
90. Van Dijk, N. M. Bounds and error bounds for queueing networks. *Annals of Operations Research*, 79: 295 319, 1998.
91. Van Doorn, E. *Stochastic Monotonicity of Birth-Death Processes*. Springer, Berlin, Lecture Notes in Statistics 4, 1981
92. Van der Wal, J. Monotonicity of the throughput of a closed exponential queueing network in the number of jobs. *OR Spectrum*, 11: 97–100, 1989.
93. Tsoucas, P. and Walrand, J. A Note on Stochastic Bounds for Queueing Networks. *Advances in Applied Probability,* 16: 926–928, 1984.
94. Tsoucas, P. and Walrand, J. Monotonicity of Throughput in Non-Markovian Networks. *Journal of Applied Probability* 26: 134–141, 1989.
95. Walrand, J. and Varaiya, P. Sojourn times and the overtaking condition in Jacksonian networks. *Advances of Applied Probability*, 12: 1000–1018, 1980.

96. Whitt, W. Uniform conditional stochastic order. *Journal of Applied Probability*, 17 : 112–123, 1980.
97. Whitt, W. Comparing counting processes and queues. *Advances in Applied Probability*, 13: 207–220, 1981.
98. Whitt, W. Uniform conditional variability ordering of probability distributions. *Journal of Applied Probability*, 22 : 619–633, 1985.
99. Whitt, W. Stochastic-Process Limits. Springer, 2002.

Chapter 9

Error Bounds and Comparison Results: The Markov Reward Approach For Queueing Networks

Nico M. Van Dijk

Abstract This chapter presents an approach to compare two Queueing Networks. Here one may typically think of one network to be a solvable modification of another únsolvable one of practical interest.

The approach is essentially based upon evaluating steady state performance measures by a cumulative reward structure and strongly relies upon the analytical estimation of so-called bias-terms. This approach, referred to as Markov Reward approach:

- may lead to (analytic) error bounds for the discrepancy
- may still apply while stochastic comparison fails

The chapter will be divided in two parts:

A General results:	which contains motivation and general results.
B Applications:	which illustrates the results and the technical verification by an instructive example and two motivational applications.

In **A** also the various advantages (as well as disadvantages) of the approach over standard stochastic comparison will be reviewed. In **B** the combination of both will be made fruitful for the truncation of Finite Jackson Networks. Some possible extensions and open questions will be addressed briefly.

Nico M. Van Dijk
University of Amsterdam, Roetersstraat 11, 1018 WB Amsterdam,
e-mail: n.m.vandijk@uva.nl

R.J. Boucherie and N.M. van Dijk (eds.), *Queueing Networks: A Fundamental Approach*,
International Series in Operations Research & Management Science 154,
DOI 10.1007/978-1-4419-6472-4_9, © Springer Science+Business Media, LLC 2011

A: General results

9.1 Motivation

9.1.1 A first example

9.1.1.1 Applications and solvability

As extension of standard one-dimensional multi-server queues Queueing Networks are widely known as a most powerful modeling tool for a variety of application such as in:

- Telephony (circuit switch networks)
- Computer networking (package switch networks)
- Manufacturing (for assembly lines or material handling systems)

but also present-day applications of rapidly growing interest as:

- Service networks (such as supply chains, call centers and hospitals)
- Mobile and ad-hoc communication networks
- and last but not least: internet

Exponential structure and solvability. Standard multi-server queueing systems, such as $M|G|c|N$ type systems, have intensively been studied under a variety of both exponential and non-exponential situations. Networks of queues, in contrast, are usually described under exponential arrival and service assumptions, as analytic results or approximations for the non-exponential case are hardly available. (Except for product form networks under special so-called insensitive service disciplines, such as pure multiserver or processor sharing disciplines.)

A queueing network (QN) model thus usually relies upon an underlying exponential structure and as such can be regarded as a continuous-time Markov chain (CTMC). The typical measures of practical interest are steady state performance measures as

- a throughput
- a mean delay
- a mean queue length
- a mean workload or efficiency
- a blocking, loss or congestion probability

Unfortunately, even for the exponential case and with the rather exclusive exception of product form type networks, closed form solutions for the steady state distribution and corresponding associated performance measures of interest as mentioned, are generally not available. These closed form, most notably product form, solutions are usually destroyed by practical features such as finite capacities, overflow, dynamic routing, breakdowns, prioritizations, synchronizations or resource contentions.

As a consequence, numerical or approximate computations will have to take place. As these are computationally expensive if not prohibitive, a number of different questions may arise. To illustrate this in more concrete form, let us first consider a simple, but yet unsolvable and instructive example.

9.1.1.2 Instructive breakdown example

Consider a simple $M|M|1|N$-system with Poisson arrival rate λ, exponential service parameter μ and a finite capacity for at most N jobs. Let n be the total number of jobs present (the job in service included). When the system is congested (i.e., $n = N$) an arrival is rejected and lost. In addition, the system is subject to breakdowns. More precisely, when the system is operative ($\theta = 1$) the system (server) can break down at an exponential rate γ_1, regardless of the number of jobs present. When the system is down ($\theta = 0$) it can become operative again, that is be repaired, at an exponential repair rate γ_0. The system status of either up ($\theta = 1$) or down ($\theta = 0$) thus follows an alternating renewal process. Let τ be the fraction of time that the system is down, i.e. $\tau = \gamma_1/(\gamma_0 + \gamma_1)$. When the system is down arrivals still take place.

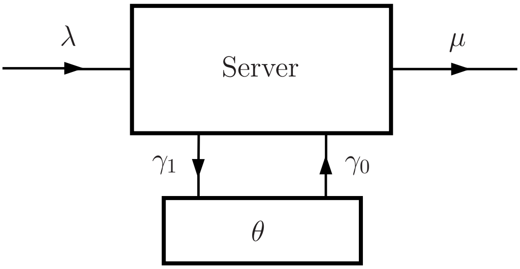


Fig. 9.1: Breakdown system.

As a first glance, this system might be seen as a most simple standard type one-dimensional queueing system. However, as both the number of jobs and the up or down status of the system is to be kept track of, it can, if not is to, be regarded as a simple network with the up-down status governed by a separate station as in figure 9.1. In fact, as simple as the system may seem, it has no simple closed form steady state distribution $\pi(n)$ for the number of jobs n present at an arbitrary instant. (Its generating function can be obtained as in Jaiswal 1968 for priority queues and recurrent relations for $\pi(n)$ can be derived). As primary measure of interest let us focus on the loss probability $\boldsymbol{B} = \pi(n = N)$ or directly relate the throughput $\boldsymbol{F} = \lambda(1 - \boldsymbol{B})$. As the system has no simple analytic expression in terms of (λ, μ, τ), numerical computation is required to compute the value of $\boldsymbol{B}$. The following questions might therefore come up:

(i)(Sensitivity error) What is the effect of an imprecision, such as due to a data estimation error, or perturbation in one of the parameters (λ, μ, τ). Can we quantify this effect analytically rather than by numerical computation.

(ii)(Approximation-Error bound) As τ must typically be thought of as being small, say in the order of a few %, wouldn't the simple $M|M|1|N$-loss probability, that is by assuming that breakdowns do not take place, lead to a reasonable approximation. Again, without numerical computation or simulation can we quantify its accuracy by an analytic error bound?

(iii)(Truncation-Error bound) As N might be quite large while $\boldsymbol{B}$ can be thought of as being of order ρ^N, for numerical reduction purposes, one might suggest to reduce the number N to a number $L << N$. Can we provide an a priori analytic quantification of the effect of this state space truncation, so as to determine a reasonable number L?

(iv)(Monotonicity-comparison result) For each of the parameters $\lambda, \mu, N, \gamma_1$ and γ_0 separately its increasing or decreasing effect on $\boldsymbol{B}$ seems obvious. But as a finite system is involved one has to be careful. At sample path basis counterintuitive examples can be constructed (e.g. see the counterintuitive example in section 9.4). Formal monotonicity proofs might thus be required. How can this be established?

(v)(Lower an upper bound-comparison result) The approximation under (ii), that is by assuming that $\tau = 0$, can intuitively be expected to give a lower bound $\boldsymbol{B}_L$. However, to guarantee a sufficiently small loss probability $\boldsymbol{B}$ (or sufficiently large throughput $\boldsymbol{F}$), such as by adjusting N, an upper bound would be of more interest. To this end, modify the system by also rejecting arrivals when the system is down. With (n, θ) representing that n jobs are present and that the system is up ($\theta = 1$) or down ($\theta = 0$), and with $1_{(A)}$ the indicator function of an event A, under this modification the global balance equation in state (n, θ) becomes:

$$
\begin{array}{lr}
\left\{\begin{array}{l}
\pi(n,\theta)\lambda 1_{(n<N)}1_{(\theta=1)} \\
\pi(n,\theta)\mu 1_{(n>0)}1_{(\theta=1)} \\
\pi(n,\theta)\gamma_1 1_{(\theta=1)} \\
\pi(n,\theta)\gamma_0 1_{(\theta=0)}
\end{array}\right\}
&
\begin{array}{l}
(9.1.1.1) \\ (9.1.1.2) \\ (9.1.1.3) \\ (9.1.1.4)
\end{array}
\\
= & \qquad\qquad (9.1.1) \\
\left\{\begin{array}{l}
\pi(n+1,\theta)\mu 1_{(n<N)}1_{(\theta=1)} \\
\pi(n-1,\theta)\lambda 1_{(n>0)}1_{(\theta=1)} \\
\pi(n,0)\gamma_0 1_{(\theta=1)} \\
\pi(n,1)\gamma_1 1_{(\theta=0)}
\end{array}\right\}
&
\begin{array}{l}
(9.1.1.1)' \\ (9.1.1.2)' \\ (9.1.1.3)' \\ (9.1.1.4)'
\end{array}
\end{array}
$$

This equation is directly verified by equating each of its four detailed equations (1.1.i)=(1.1.i)$'$ for $i = 1,2,3,4$ separately, by substituting the product form, with c a normalizing constant,

$$\pi(n,\theta) = c[\gamma_\theta]^{-1}\left[\frac{\lambda}{\mu}\right]^n \quad (n \le N) \tag{9.1.2}$$

Intuitively, we can now expect to obtain an upper bound $\boldsymbol{B}_U = \pi(n = N, \theta = 1) + \pi(\theta = 0)$. Hence, it seems appealing to conclude that

$$\boldsymbol{B}_L \le \boldsymbol{B} \le \boldsymbol{B}_U \tag{9.1.3}$$

Here $\boldsymbol{B}_L$ would be obtained by the standard $M|M|1|N$-loss probability by assuming that the system never breaks down (i.e.: $\tau = 0$). The inequalities (9.1.3) as well as their practical usefulness are also supported numerically in table 9.1 below.

Table 9.1: Lower and upper bounds for the loss fraction $\boldsymbol{B}$

N	ρ	τ	$\boldsymbol{B}_L$	$\boldsymbol{B}_U$
20	20	0.1	0.16	0.24
		0.05	0.16	0.20
		0.02	0.16	0.18
30	25	0.05	0.052	0.098
		0.01	0.052	0.062

Nevertheless, as shown in section 9.4, at sample path basis one can provide counterintuitive examples by which an ordering as in (9.1.3) seems violated. It thus seems of both practical an theoretical interest to formally prove the bounds (9.1.3).

9.1.1.3 Two questions

In essence, each of the questions for the instructive example comes down to the comparison of two systems. One which can be regarded as an original system and one as a modified one, say as due to:

- a perturbation of an input parameter,
- a system modification,
- or a state space truncation

Particulary the situation of a system modification can be of considerable practical interest so as to justify a computational simplification, by either:

- an analytic error bound for its accuracy, or
- a secure bound for the performance measure of interest.

Here one may typically (but not necessarily) think of system modifications that will lead to a product form computation. Accordingly, the two major questions of interest when comparing two related systems, say an original and a modified one, become:

Q1: How to obtain comparison or ordering results, that is with $\geq$ or $\leq$ sign, so as to guarantee bounds.

Q2: How to obtain analytic error bounds on the discrepancy of the two systems, for some specific performance measure.

Here the order of the two questions is interchanged from its practical motivation as the first seems more standard and easier to handle first (as will also appear later on).

Before stepping into more detail of the objective and the approach of how to address these two questions, let us give two more motivating network examples for each of these two questions, which are of practical interest by itself. Also these examples will be dealt with later on.

9.1.2 Two more examples

9.1.2.1 Finite Tandem Queues

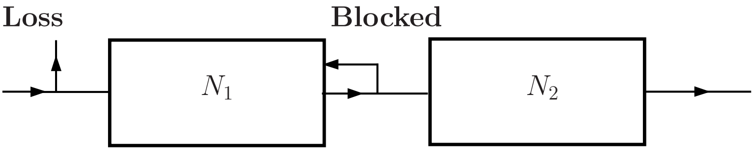


Fig. 9.2: Tandem system.

Consider a two-station tandem system with capacity constraints for at most N_1 jobs at station 1 and N_2 jobs at station 2. When station 1 is saturated, arrivals are rejected and lost. When station 2 is saturated, the servicing at station 1 is stopped. This system can be regarded as representative for a variety of applications in manufacturing (assembly lines) and computer performance evaluation (multi-stage processing). Due to the finite constraints, however, it has **no** product-form expression. Various numerical and approximation techniques have therefore been developed (e.g. [36], [7], [10], [30]). These, however, still require restrictive service specifications (such as exponential), are computationally expensive, ánd last but not least, do nót provide any guarantee or error bound.

In order to enforce a product form expression, the following two modifications can now be suggested.

Modification 1:

Never stop station 1. Reject arrivals only when $n_1 + n_2 = N_1 + N_2$.

Modification 2:

- When the second station is saturated also reject arrivals at station 1.
- When the first station is saturated also stop (the servicing at) station 2.

Let λ be the arrival rate and μ_i the exponential service parameter, assuming a single server, at station $i = 1,2$. Then indeed, with n_i the number of jobs at station i, $i = 1,2$, for modification 2 one easily verifies the global balance equation (9.1.4) by equating each of the detailed equation (1.4.i)=(1.4.i)$'$ for $i = 1,2,3$.

$$\begin{array}{ll}
\left\{\begin{array}{l} \pi(n_1,n_2)\lambda 1_{(n_1<N_1)} 1_{(n_2<N_2)} \\ \pi(n_1,n_2)\mu_1 1_{(n_1>0)} 1_{(n_2<N_2)} \\ \pi(n_1,n_2)\mu_2 1_{(n_2>0)} 1_{(n_1<N_1)} \end{array}\right\} & \begin{array}{l}(9.1.4.1)\\(9.1.4.2)\\(9.1.4.3)\end{array} \\
= & \\
\left\{\begin{array}{l} \pi(n_1,n_2+1)\mu_2 1_{(n_1<N_1)} 1_{(n_2<N_2)} \\ \pi(n_1-1,n_2)\lambda 1_{(n_1>0)} 1_{(n_2<N_2)} \\ \pi(n_1+1,n_2-1)\mu_1 1_{(n_2>0)} 1_{(n_1<N_1)} \end{array}\right\} & \begin{array}{l}(9.1.4.1)'\\(9.1.4.2)'\\(9.1.4.3)'\end{array}
\end{array} \tag{9.1.4}$$

by substituting the product form:

$$\pi(n_1,n_2) = c\left(\frac{\lambda}{\mu_1}\right)^{n_1}\left(\frac{\lambda}{\mu_2}\right)^{n_2} \qquad , 0 \le n_1 \le N_1 \; ; \; 0 \le n_2 \le N_2 \; ; \quad n_1 + n_2 \ne N_1 + N_2 \; .$$

In a similar fashion the same product form is also verified under modification 1 but for all states with $n_1 + n_2 \le N_1 + N_2$. Now suppose again, that we are interested in the blocking probability $\boldsymbol{B}$ for the original tandem queue.

Intuitively, modification 1 will lead to a lower bound $\boldsymbol{B}_L$ and modification 2 to an upper bound $\boldsymbol{B}_U$. That is, inequality (9.1.3) can again be expected intuitively. This is also supported by some numerical results in table 9.2. These results also indicate a practical usefulness. Nevertheless, again one can construct counterintuitive examples at sample path basis that seem to conflict with (9.1.3). Given the generic nature of this example, particularly in this case a formal proof for the bounds in (9.1.3) would thus be of both theoretical and considerable practical interest.

9.1.2.2 Finite Jackson Networks

The famous class of so-called Jackson networks, named after the pioneering paper by Jackson in 1957, refers to networks which allow a random routing of jobs from one service station to another with fixed probabilities. This subclass forms a rich

Table 9.2: Comparison of Bounds and Numerical Results

N_1	N_2	λ	μ_1	μ_2	$\boldsymbol{B_L}$	$\boldsymbol{B}$	$\boldsymbol{B_U}$
10	10	1	1	1	0.090	0.124	0.167
20	20	1	1	1	0.047	0.064	0.091
40	20	1	1	1	0.003	0.003	0.070

class for practical applications. The steady state distribution of the standard Jackson network, that is with infinite capacities, is well known to exhibit an appealing product form. By this product form expression performance measures of interest, such as mean delays, mean queue lengths or throughputs of service stations can be computed directly.

In practice, however, also finite capacity constraints on the numbers of jobs at these service stations are most natural, say for at most N_i jobs at station $i = 1,\dots,J$ with J the number of stations. Unfortunately, these finite constraints generally violate the product form (or any closed form) expression. Next to approximate methods and simulation, a numerical computation of system performance measures thus becomes of practical interest.

However, due to the multi-dimensional structure of the queueing network, the size of the corresponding state space can be large if not astronomic. The computational effort therefore will rapidly become expensive if not prohibitive. A reduction of the state space by reducing the numbers N_i to $L_i \leq N_i$ might thus become appealing. Intuitively, as networks are developed such that congestion probabilities are small, the effect of these truncations can still be small. Clearly, such state space reductions will have been applied frequently in practice. Nevertheless, formal theoretical support in terms of error bounds, either at computational basis itself or in analytic form, seems to be lacking. Alternatively, one might wish to expand the numbers N_i to $N_i = \infty$ so as to regain a product form. For either case, a truncation or an expansion, an analytic error bound for the effect of the modification will thus be of interest.

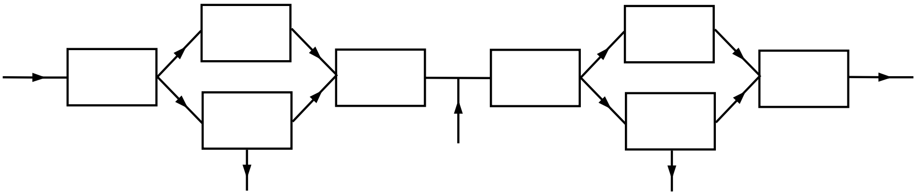


Fig. 9.3: Jackson network.

9.1.3 Objectives

As motivated by sections 9.1.1 and 9.1.2, the objectives of this chapter are:

(i) To show that the two questions of interest: **Q1** and **Q2** from section 9.1.1.3, can be addressed in a unified manner.

(ii) To provide a separate comparison, error bound and truncation result.

(iii) To illustrate the conditions required, the verification of these conditions and the different type of comparison and error bounds results that can be obtained.

9.1.4 Approach

To this end, in contrast with the standard stochastic comparison approach, a Markov reward approach will be presented. This approach is based upon a discrete-time transformation and one-step Markov reward or dynamic programming steps.

In essence, in its discrete-time formulation, this approach is strongly related to the policy improvement step in classical stochastic dynamic programming (e.g. [52], [50]) and as such could be regarded as well-known. Nevertheless, for the two questions **Q1** and **Q2** mentioned, to the best of knowledge, the approach has been proposed and been applied first in [19] for question **Q1** and [23] for question **Q2**.

More detailed, the essential ingredients of this approach are:

- To analyze steady state performance measures as by expected average rewards.
- To use a discrete-time Markov transition structure and to compare the difference of the two systems in its one-step transition structure.
- To use inductive arguments to estimate or bound so-called bias (or relative gain) terms for one of the two systems.

For comparison results (that is, question **Q1**) the ingredients of the MRA will lead to both some disadvantages and some advantages as opposed to the standard stochastic comparison approach, as developed most elegantly in the book by [61] and related references thereafter (e.g. [43], [47], [57], [63]). (Advantages and disadvantages of the MRA will be listed in more detail in section 9.3.5). As a major disadvantage the MRA requires a Markovian and thus exponential Queueing Network structure (although possible extensions to non-exponential situations will be mentioned briefly later on). As an advantage, however, the MRA might work for some specific performance measure while stochastic comparison might not. (As will be illustrated in sections 9.2.2 and 9.5).

9.1.5 Outline

First, in section 9.2, some notation and preliminary results will be presented to transform an exponential queueing in a discrete time Markov Chain. Next, a basic result for stochastic comparison of two systems is presented but also shown to be limited for specific applications such as the comparison for the finite tandem queue from section 9.1.2.1.

Next, in section 9.3 therefore, the Markov reward approach (MRA) is introduced to overcome this limitation as well as to provide analytic error bounds for the discrepancy of two systems. The entire section is set up in the general context of continuous-time Markov chains, with queueing networks as a special but primary motivational application in mind.

In section 9.3.1 the necessary framework of a one-step reward structure will be set up. Next, in section 9.3.2 the application of the Markov reward approach for the comparison of two systems is presented (result 9.3.2). In section 9.3.3, as a main result of the approach, an error bound result is developed (result 9.3.10). In section 9.3.4 the error bound result is also made more explicit for a state space truncation of a Markov chain, such as a queueing network. Though the results in sections 9.3.2, 9.3.3 and 9.3.4 are strongly related in form, proofs and technical verification of its conditions, the results are given separately to contrast more explicitly with the stochastic comparison approach as well as to highlight the specific aspects for the different type of applications. Finally, in section 9.3.5, the major advantages as well as disadvantages of the Markov reward approach as opposed to stochastic comparison approach are briefly mentioned.

The application of these general results to queueing networks the verification of the conditions and the possible concrete results will then be illustrated in sections 9.4, 9.5 and 9.6.

- In section 9.4 for the instructive breakdown example of section 9.1.1.2.
- In section 9.5 for the finite tandem queue bounds from section 9.1.2.1.
- In section 9.6 for the finite Jackson network of section 9.1.2.2.

First, in section 9.4 the instructive breakdown example from section 9.1.1.2 will be dealt with to illustrate the verification of the conditions and the results from section 9.3. Particularly, it will be shown how the crucial step (the bounding of so-called bias-terms) can be achieved in an analytic and unifying manner for different performance measures. This step will be worked out in detail so that the equations that come along can also be understood more easily in more complex situations, as in sections 9.5 and 9.6. Next, in section 9.5, the motivational example of a finite tandem queue will be considered to prove a simple lower and upper performance bound.

Finally, in section 9.6, the truncation is investigated of Finite Jackson Networks for computational simplification. Both a computational and analytic error bound are derived. For the example of a cellular mobile network an analytic relative error bound is established as based upon standard infinite server queues.

A brief discussion and evaluation in section 9.7 concludes the chapter. This includes possible extensions to non-exponential queueing networks, to the case of transient measures, to discrete-time queueing networks or more general continuous-time systems governed by nonnegative matrices, as well as open questions for further research and some other recent applications of the approach.

9.2 Stochastic Comparison.

9.2.1 Preliminaries

As we will restrict to exponential queueing networks, throughout section 9.2 and 9.3 we will consider continuous-time Markov chains (CTMC) with countable state space S and transition rate matrix $\mathbf{Q} = \boldsymbol{q}(i,j)$, with $\boldsymbol{q}(i,j)$ the transition rate for a change from state i into state $j \neq i$ and $\boldsymbol{q}(i,i) = -\sum_{j\neq i}\boldsymbol{q}(i,j)$. For convenience, this chain is assumed to be *uniformizable*. That is, for some finite constant $H < \infty$ and all $i \in \boldsymbol{S}$,

$$\sum_{j\neq i}\boldsymbol{q}(i,j) \leq H \tag{9.2.1}$$

Let $\boldsymbol{P}_t(i,j)$ denote the transition probability for a transition from state i into state j over time t and define expectation operators $\{\boldsymbol{T}_t \mid t \geq 0\}$ on the set $\mathbb{B}$ of real-valued functions f defined on $\boldsymbol{S}$ by

$$(\boldsymbol{T}_t f)(i) = \sum_j \boldsymbol{P}_t(i,j)f(j). \tag{9.2.2}$$

In words that is, $(\boldsymbol{T}_t f)(i)$ represents the expected value of function f at time t of the CTMC when starting in state i at time 0. By virtue of the boundedness (uniformization) assumption (9.2.1), it is then well known (e.g. [24], [31], [33]) that the continuous-time Markov chain can also be evaluated as a discrete-time Markov chain (DTMC) with one-step transition matrix $\boldsymbol{P}$ with $h = H^{-1}$:

$$\boldsymbol{P} = \mathbf{I} + h\boldsymbol{Q},$$

hence, with one-step transition probabilities:

$$\boldsymbol{P}(i,j) = \begin{cases} h\boldsymbol{q}(i,j) & (j \neq i), \\ 1 - h\sum_{j\neq i}\boldsymbol{q}(i,j) & (j = i). \end{cases} \tag{9.2.3}$$

Intuitively speaking, one may regard this matrix as a transition matrix over a time interval of length $h = H^{-1}$. In contrast with the CTMC, however, it ignores possible multiple changes in this time interval. Nevertheless it can be shown that the stochastic behavior of the CTMC, more precisely, the transition mechanisms and corresponding expectation over any time t, can stochastically be obtained as if at exponential times with parameter H, thus on average per time interval of length $h = H^{-1}$, a change may take place as according to the one-step transition matrix $\boldsymbol{P}$. Let $\boldsymbol{T}^k$ for the DTMC represent (similar to $\boldsymbol{T}_t$ for the CTMC) the expectation operator over k steps (here $\boldsymbol{P}^k$ denotes the k-th matrix power of $\boldsymbol{P}$ and $\boldsymbol{I}$ is the identity operator), i.e.:

$$\begin{cases} \boldsymbol{T}^0 f(i) := f(i) \\ \boldsymbol{T}f(i) \ := \sum_j \boldsymbol{P}(i,j)f(j) \\ \boldsymbol{T}^k f(i) := \sum_j \boldsymbol{P}^k(i,j)f(j) = \boldsymbol{T}(\boldsymbol{T}^{k-1}f)(i) \ (k>0, \text{ for all } f \in \mathbb{B}). \end{cases} \tag{9.2.4}$$

Then, under natural ergodicity and irreducibility conditions we may conclude that for the steady-state performance measure of interest $\boldsymbol{G}$ and independent of $i \in \boldsymbol{S}$:

$$\boldsymbol{G} = \lim_{t\to\infty} \boldsymbol{T}_t r(i) \quad \text{and} \quad \boldsymbol{G} \overset{C}{=} \lim_{k\to\infty} \boldsymbol{T}^k r(i) = \lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} \boldsymbol{T}^k r(i) \tag{9.2.5}$$

(where for the discrete-time case the Cesaro limit is used to cover aperiodicity), for some appropriate reward rate r. $\boldsymbol{G}$ is thus be regarded as a scalar which represents the expected average reward per unit time in steady state situation. For example, with the CTMC representing an $M|M|1|N$ queue: and n the number of jobs present

$$\boldsymbol{G} = \begin{cases} \text{Mean queue length} & \text{for } r(n) = n \\ \text{Loss probability} & \text{for } r(n) = 1_{(n=N)} \\ \text{Throughput} & \text{for } r(n) = \mu 1_{(n>0)} \end{cases} \tag{9.2.6}$$

9.2.2 Stochastic comparison

Now suppose that we like to compare a performance measure

$$\begin{cases} \boldsymbol{G} \text{ for an original CTMC with transition rates } \boldsymbol{q}(i,j) \text{ with} \\ \bar{\boldsymbol{G}} \text{ for a modified CTMC with transition rates } \bar{\boldsymbol{q}}(i,j) \end{cases}$$

say at one and the same state space $\boldsymbol{S} = \bar{\boldsymbol{S}}$, for both of which the uniformization condition (9.2.1) holds with same constant H, and where $\boldsymbol{G}$ and $\bar{\boldsymbol{G}}$ are the average expected rewards for one the same reward rate r.

Let $\boldsymbol{T}_t$ and $\bar{\boldsymbol{T}}_t$, $\boldsymbol{T}$ and $\bar{\boldsymbol{T}}$ as well as $\boldsymbol{T}^k$ and $\bar{\boldsymbol{T}}^k$ be the corresponding uniformized one-step transition operators as defined by (9.2.2), (9.2.3) and (9.2.4). Then by virtue of the uniformization, that is (9.2.5), we can use

$$\begin{aligned} \boldsymbol{G} &\overset{C}{=} \lim_{k\to\infty} \boldsymbol{T}^k r(i) \quad &&,\text{for any } i \in \boldsymbol{S} \\ \bar{\boldsymbol{G}} &\overset{C}{=} \lim_{k\to\infty} \bar{\boldsymbol{T}}^k r(i) \quad &&,\text{for any } i \in \boldsymbol{S} \end{aligned} \tag{9.2.7}$$

The following (strong) stochastic monotonicity or comparison result can now be concluded directly for comparing $\boldsymbol{G}$ and $\bar{\boldsymbol{G}}$. Herein, for two functions f and g we write $f \geq g$ iff $f(i) \geq g(i)$ for all $i \in \boldsymbol{S}$. Furthermore, the function 0 represents the null function, i.e. $0(i) = 0$ for any $i \in \boldsymbol{S}$.

Result 9.2.1 (Stochastic monotonicity) *Let $\mathbb{M}$ represent a set of real-valued (monotonicity) functions which is closed under $\boldsymbol{T}$, i.e.*

$$\boldsymbol{T}f \in \mathbb{M} \quad \textit{for any} \quad f \in \mathbb{M} \tag{9.2.8}$$

If

$$\bar{\boldsymbol{T}}f \geq \boldsymbol{T}f \qquad (f \in \mathbb{M}) \tag{9.2.9}$$

and

$$r \in \mathbb{M} \tag{9.2.10}$$

then

$$\bar{\boldsymbol{G}} \geq \boldsymbol{G} \tag{9.2.11}$$

Proof. By (9.2.4) for arbitrary f, any state i and $k > 0$, we can write:

$$\begin{aligned} &(\bar{\boldsymbol{T}}^k f - \boldsymbol{T}^k f)(i) = \\ &(\bar{\boldsymbol{T}}\bar{\boldsymbol{T}}^{k-1} f - \boldsymbol{T}\boldsymbol{T}^{k-1} f)(i) = \\ &(\bar{\boldsymbol{T}} - \boldsymbol{T})\boldsymbol{T}^{k-1} f(i) + \bar{\boldsymbol{T}}(\bar{\boldsymbol{T}}^{k-1} - \boldsymbol{T}^{k-1}) f(i) = \\ &\sum_{t=0}^{k-1} \bar{\boldsymbol{T}}^t \left[(\bar{\boldsymbol{T}} - \boldsymbol{T})\boldsymbol{T}^{k-t-1} f\right](i) + \bar{\boldsymbol{T}}^k \left[(\bar{\boldsymbol{T}}^0 - \boldsymbol{T}^0) f\right](i) \end{aligned} \tag{9.2.12}$$

First note that $(\bar{\boldsymbol{T}}^0 - \boldsymbol{T}^0) f(j) = f(j) - f(j) = 0$ for any j. The last term in (9.2.12) can thus be deleted. Next, note that by repetition of condition (9.2.8):

$$\boldsymbol{T}^s f \in \mathbb{M} \text{ for any } f \in \mathbb{M} \text{ and } s = k - t - 1 \geq 0 \tag{9.2.13}$$

Hence, by condition (9.2.9), for any $t \leq k - 1$ and any $f \in \mathbb{M}$:

$$(\bar{\boldsymbol{T}} - \boldsymbol{T})\boldsymbol{T}^{k-t-1} f \geq 0 \tag{9.2.14}$$

By also noting that $\bar{\boldsymbol{T}}^t$ is a (probability) transition (and thus a non-negative) operator so that for any function $g \geq 0$ (in componentwise sense) $\bar{\boldsymbol{T}}^t g \geq 0$, the right hand side of (9.2.12) can be estimated from below by 0. Condition (9.2.10) and relation (9.2.7) complete the proof. □

Remark 9.2.2 *Clearly, result 9.2.1 remains identical with reversed signs $\leq$. Result 9.2.1 is strong as it secures an ordering (a comparison) of the performance measure for all possible reward rate functions $r \in \mathbb{M}$. In addition, as shown in various references as [9], [10], [43], [38], [39], [51], [47], [61], [63], [68], [69], relaxations and extensions of this form of monotonicity results can be provided, most notably among which to non-exponential situations. In this respect it also mentioned that these stochastic comparison results also relate to the approach of stochastic (or weak) coupling and sample path comparison, by which exponentiality assumptions can be disregarded directly. But in essence, also under these relaxations and extensions, and despite its elegancy and non-exponential advantage, stochastic comparison, as reflected by result 9.2.1, can be insufficient in two ways:*

*(i) The ordering (9.2.11) holds for any $r \in \mathbb{M}$. But as a price to pay also the conditions (9.2.8) and (9.2.9) should be satisfied for any $f \in \mathbb{M}$. At this point the monotonicity class $\mathbb{M}$ is not specified. In fact, no such class may exist that also covers condition (9.2.10) for the performance measure and thus reward rate of interest: r. Differently said, the specific performance measure of interest and transition structure may require monotonicity that is nót preserved under **T**. This will be illustrated below in section 9.2.3 for the tandem example from section 9.1.2.1. Nevertheless, for (some or) a specific reward rate(s) r one might still expect an ordering result as in (9.2.11) but not for a complete closed class of functions $\mathbb{M}$. This is thus to be proved by a different type of approach. For the same tandem example, this will be shown in section 9.5.*

(ii) The comparison result nor its proof seem to lead to a quantification of the discrepancy (i.e.: an error bound) between the two systems, say as due to a perturbation or modification. For example, as a natural perturbation, assume that

$$||(\bar{\boldsymbol{T}}-\boldsymbol{T})f|| \leq \varepsilon||f|| \quad ,(f \in \mathbb{M})$$

in usual supremum norms for some small $\varepsilon > 0$. Since $\bar{\boldsymbol{T}}$ is a probability operator: $||\bar{\boldsymbol{T}}^k g|| \leq ||g||$ for any function g. As a consequence, by expansion (9.2.12) (or alternatively its one step recursion relation (9.2.4) and induction) we easily prove:

$$||\bar{\boldsymbol{T}}^k f - \boldsymbol{T}^k f|| \leq \varepsilon k||f||$$

*However, due to the limits in (9.2.7) there is **nó** such result for:*

$$|\bar{\boldsymbol{G}} - \boldsymbol{G}|$$

.

9.2.3 Stochastic comparison failure

The stochastic comparison approach has shown to be most appealing and useful for $M|G|c$-type systems and variations thereupon ([27], [59], [60], [61]); roughly speaking that is, for one-dimensional systems with a single service station or, as in [43] multi-component reliability systems, which in essence can be regarded as a machine repair or Engset type queueing system. With multiple service stations, however, as in queueing networks, its application seems less common. Some exceptions here are found in [1], [2], [53], [66], [70]. These, however, generally concern Jackson type networks without capacity constraints or other conflicting features by which service stations become directly dependent. To get more insight in the complications that arise for stochastic comparison, let us reconsider the finite tandem example from section 9.1.2.

Finite tandem example (section 9.1.2) revisited. Consider the original finite tandem queue as well as the modified model under modification 2, as described in section 9.1.2. Let $\boldsymbol{P}$ and $\bar{\boldsymbol{P}}$ be the corresponding one-step transition matrices for the uniformized DTMC, where condition (9.2.1) is guaranteed by $H = [\lambda + \mu_1 + \mu_2]$, as by (9.2.3) and (9.2.4) with a state i identified as a state (n_1, n_2) for the number of jobs n_i at station $i = 1,2$. More precisely, for the original finite tandem queue the uniformized transition matrix $\boldsymbol{P}$ becomes:

$$\boldsymbol{P}\left((n_1,n_2),(n_1,n_2)'\right) = \begin{cases} & \underline{(n_1,n_2)'} \\ 1 - h\lambda 1_{(n_1<N_1)} - h\mu_1 1_{(n_2<N_2)} & (n_1,n_2) \\ h\lambda 1_{(n_1<N_1)} & (n_1+1,n_2) \\ h\mu_1 1_{(n_2<N_2)} 1_{(n_1>0)} & (n_1-1,n_2+1) \\ h\mu_2 1_{(n_2>0)} & (n_1,n_2-1) \end{cases} \tag{9.2.15}$$

and similarly for the modified system. In order to prove that the modified (product form) system leads to an upper bound for the loss probability $\boldsymbol{B}$, we would like to apply result 9.2.1. However, the loss probability will also require a different reward rate r and $\bar{r}$ for the two systems, which is not allowed by result 9.2.1. To avoid this minor complication we can also analyze the throughput by:

$$\boldsymbol{F} = \lambda(1-\boldsymbol{B}) \quad \text{and} \quad r(n_1,n_2) = \mu_2 1_{(n_2>0)}.$$

Note that r is nondecreasing in n_2 (as well as in n_1). By comparing the transition structures for the original and modified (under modification 2) tandem queue, for arbitrary function f we conclude

$$\begin{aligned} &(\bar{\boldsymbol{T}} - \boldsymbol{T})f(n_1,n_2) = \\ &\quad h\lambda 1(n_1 < N_1, n_2 = N_2)[f(n_1,N_2) - f(n_1+1,N_2)] + \\ &\quad h\mu_2 1(n_2 > 0, n_1 = N_1)[f(N_1,n_2) - f(N_1,n_2-1)] \end{aligned} \tag{9.2.16}$$

In order to estimate this expression from above by 0 (note that the throughput of the original should be proven to be larger than for the modified system, i.e. $\bar{\boldsymbol{F}} \leq \boldsymbol{F}$), the monotonicity class $\mathbb{M}$ should thus be of the form:

$$\mathbb{M} = \{f : \boldsymbol{S} \to \mathbb{R} \mid f(n_1+1,n_2) - f(n_1,n_2) \geq 0; \\ f(n_1,n_2+1) - f(n_1,n_2) \leq 0, \text{ all } (n_1,n_2)\} \quad (9.2.17)$$

But then we directly observe that condition (9.2.10) fails as $r \notin \mathbb{M}$. In fact, also condition (9.2.8) is violated.

More precisely, to satisfy (9.2.8), these monotonicities should also apply for the function $(\boldsymbol{T}f)$ for any such f. However, by (9.2.15) for any state (n_1,n_2) and (n_1+1,n_2) with $n_1+1 \leq N_1$:

$$\begin{aligned}
&(\boldsymbol{T}f)(n_1+1,n_2) - (\boldsymbol{T}f)(n_1,n_2) \\
&\quad = h\lambda 1_{(n_1+1<N_1)}[f(n_1+2,n_2) - f(n_1+1,n_2)] \\
&\quad + h\lambda 1_{(n_1+1=N_1)}[f(n_1+1,n_2) - f(n_1+1,n_2)] \\
&\quad + h\mu_1 1_{(n_1>0)} 1_{(n_2<N_2)}[f(n_1,n_2+1) - f(n_1-1,n_2+1)] \\
&\quad + h\mu_1 1_{(n_1=0)} 1_{(n_2<N_2)}[f(n_1,n_2+1) - f(n_1,n_2)] \\
&\quad + h\mu_2 1_{(n_2>0)}[f(n_1+1,n_2-1) - f(n_1,n_2-1)] \\
&\quad + [1 - h\lambda 1_{(n_1+1<N_1)} - h\lambda 1_{(n_1+1=N_1)} - h\mu_1 1_{(n_2<N_2)} - h\mu_2 1_{(n_2>0)}] \\
&\qquad [f(n_1+1,n_2) - f(n_1,n_2)]
\end{aligned} \quad (9.2.18)$$

To conclude that this expression is larger than or equal to 0 we can use the monotonicity of f in n_1 in the first, (the second is equal to 0 itself), third, fifth and sixth term in the right hand side. However, in a state with $n_1 = 0$ and $n_2 < N_2$, there is a serious conflict in the fourth term as $f(n_1,n_2+1) - f(n_1,n_2) \leq 0$, as by (9.2.17). In other words, with $\mathbb{M}$ as by (9.2.17) condition (9.2.8) fails.

Would we have defined $\mathbb{M}$ with functions that are nondecreasing in both components, i.e.

$$\mathbb{M} = \{f : \boldsymbol{S} \to \mathbb{R} \mid f(n_1+1,n_2) \geq f(n_1,n_2); \\ f(n_1,n_2+1) \geq f(n_1,n_2), \text{ all } (n_1,n_2)\} \quad (9.2.19)$$

(9.2.18) could indeed be estimated from below by 0. And also for the second component, as by

$$\begin{aligned}
&(\boldsymbol{T}f)(n_1,n_2+1)-(\boldsymbol{T}f)(n_1,n_2)\\
&\quad = h\lambda 1_{(n_1<N_1)}[f(n_1+1,n_2+1)-f(n_1+1,n_2)]\\
&\quad + h\mu_1 1_{(n_1>0)} 1_{(n_2+1<N_2)}[f(n_1-1,n_2+2)-f(n_1-1,n_2+1)]\\
&\quad + h\mu_1 1_{(n_1>0)} 1_{(n_2+1=N_2)}[f(n_1,n_2+1)-f(n_1-1,n_2+1)]\\
&\quad + h\mu_2 1_{(n_2>0)}[f(n_1,n_2)-f(n_1,n_2-1)]\\
&\quad + h\mu_2 1_{(n_2=0)}[f(n_1,n_2)-f(n_1,n_2)]\\
&\quad + [1-h\lambda 1_{(n_1<N_1)}-h\mu_1 1_{(n_1>0)}-h\mu_2][f(n_1,n_2+1)-f(n_1,n_2)]
\end{aligned} \tag{9.2.20}$$

we can conclude that the monotonicity is preserved. In other words, with $\mathbb{M}$ as by (9.2.20) condition (9.2.8) is satisfied. In addition, we also have:

$$r \in \mathbb{M}$$

However, while (9.2.8) and (9.2.10) are now satisfied, to apply result 9.2.1, the ordering condition (9.2.9) will necessarily fail, as due to (9.2.16).

Conclusion. *In other words, there seems no way to apply the stochastic comparison result 9.2.1 to prove the upper bound* $\boldsymbol{B}_U$ *in section 9.1.2.1 (or, by similar counter arguments, the lower bound* $\boldsymbol{B}_L$*).*

This is nót to say that stochastic comparison results cannot be obtained for the finite tandem queue (as argued above, it will apply for monotone functions of the form (9.2.17), but nót for the specific performance measure and system to be compared with, as of interest by its motivation in section 9.1.2.1.

9.3 Markov reward approach

9.3.1 Preliminaries

As argued in remark 9.2.2(ii), despite the fact that an average performance measure can be regarded as an expectation at an arbitrary instant, it seems impossible to conclude an error bound by just analyzing the effect on expectations at finite instants. We will therefore use a cumulative reward structure.

Consider some given reward rate function $r(i)$ that incurs a reward $r(i)$ per unit of time whenever the system is in state i. The expected cumulative reward over a period of length t and given the initial state i at time 0 is then given by

$$\boldsymbol{V}_t(i) = \int_0^t \boldsymbol{T}_s r(i) ds. \tag{9.3.1}$$

Then, as in (9.2.5), under natural ergodicity conditions this expected cumulative reward averaged over time will converge to the expected average reward, or in the current setting, the performance measure $\boldsymbol{G}$, independently of the initial state i, as:

$$\boldsymbol{G} = \lim_{t\to\infty} \frac{1}{t} \boldsymbol{V}_t(i) \qquad (\text{for any } i \in \boldsymbol{S}) \tag{9.3.2}$$

By virtue of the uniformization technique again, we can also evaluate $\boldsymbol{G}$ by means of the expected cumulative reward for the uniformized discrete time Markov chain as:

$$\boldsymbol{G} = \lim_{k\to\infty} \frac{H}{k} \boldsymbol{V}^k(i) \qquad (\text{for any } i \in \boldsymbol{S}) \tag{9.3.3}$$

Here $\boldsymbol{V}^k(i)$ represents the expected cumulative reward for the uniformized DTMC over k steps, each of length $h = H^{-1}$, with one-step rewards $hr(j)$ per step whenever the system is in state j and when starting in state i at time 0. More precisely, for any $i \in S$

$$\boldsymbol{V}^k(i) = \sum_{s=0}^{k-1} h\boldsymbol{T}^s r(i), \quad (k = 1,2,\ldots), \quad \boldsymbol{V}^0(i) = 0 \quad (i \in \boldsymbol{S}) \tag{9.3.4}$$

The factor H in (9.3.3) is required as the time average of $\boldsymbol{V}^k/k$ ensures an average reward per step of mean length $h = H^{-1}$ instead of per unit of time.

The major advantage of this discrete setup is that it enables one to use **inductive** arguments by exploiting the reward (or dynamic programming) relation:

$$\boldsymbol{V}^{k+1}(i) = hr(i) + \sum_j \boldsymbol{P}(i,j)\boldsymbol{V}^k(j) \qquad (k = 0,1,2,\ldots),(i \in \boldsymbol{S}) \tag{9.3.5}$$

Remark 9.3.1 (Scaling factors) *Clearly, the time scaling factors h and H in (9.3.3), (9.3.4) and (9.3.5) could be deleted. However, they are left in for their natural interpretation accordingly to the uniformization, as will also appear to be convenient for the probabilistic interpretation of the so-called bias-terms equations as will be derived later on (in sections 9.4, 9.5, 9.6).*

9.3.2 Comparison Result

Consider a CTMC, which will be referred to as original model, as described in section 9.2.1, with transition rates $\boldsymbol{q}(i,j)$, reward rate $r(i)$ and state space $\boldsymbol{S}$. We briefly denote this parametrization by $(\boldsymbol{S},\boldsymbol{q},r)$. Now consider a second CTMC, described similarly, which will be thought of and be referred to as a modified model of the first, $(\bar{\boldsymbol{S}},\bar{\boldsymbol{q}},\bar{r})$. In short, we aim to compare these two systems:

$$\begin{cases} (\boldsymbol{S}, \boldsymbol{q}, r) \\ (\bar{\boldsymbol{S}}, \bar{\boldsymbol{q}}, \bar{r}) \end{cases} \text{under the condition } \bar{\boldsymbol{S}} \subseteq \boldsymbol{S}. \tag{9.3.6}$$

Here both models are assumed to be unifomizable with same constant H as by (9.2.1). Throughout, we use the overbar symbol for an expression concerning the modified model. We aim to compare the performance measures, that is the expected reward per unit time in steady state, $\boldsymbol{G}$ and $\bar{\boldsymbol{G}}$.

Result 9.3.2 *Suppose that for $\boldsymbol{V}^k$ defined by (9.3.5), all $i \in \bar{\boldsymbol{S}}$ and $k \geq 0$:*

$$[\bar{r} - r](i) + \sum_j [\bar{\boldsymbol{q}}(i,j) - \boldsymbol{q}(i,j)][\boldsymbol{V}^k(j) - \boldsymbol{V}^k(i)] \geq 0 \tag{9.3.7}$$

Then,

$$\bar{\boldsymbol{G}} \geq \boldsymbol{G}. \tag{9.3.8}$$

Proof. By virtue of (9.3.5), we have

$$\begin{aligned} \boldsymbol{V}^{k+1}(i) &= hr(i) + \boldsymbol{T}\boldsymbol{V}^k(i), \\ \bar{\boldsymbol{V}}^{k+1}(i) &= h\bar{r}(i) + \bar{\boldsymbol{T}}\bar{\boldsymbol{V}}^k(i), \end{aligned} \tag{9.3.9}$$

As the transition probabilities $\bar{\boldsymbol{P}}(\cdot,\cdot)$ remain restricted to $\bar{\boldsymbol{S}} \subseteq \boldsymbol{S}$, for arbitrary $l \in \boldsymbol{S}$ we may thus write

$$\begin{aligned} &(\bar{\boldsymbol{V}}^k - \boldsymbol{V}^k)(l) \\ &\quad = h(\bar{r} - r)(l) + (\bar{\boldsymbol{T}}\bar{\boldsymbol{V}}^{k-1} - \boldsymbol{T}\boldsymbol{V}^{k-1})(l) \\ &\quad = h(\bar{r} - r)(l) + (\bar{\boldsymbol{T}} - \boldsymbol{T})\boldsymbol{V}^{k-1}(l) + \bar{\boldsymbol{T}}(\bar{\boldsymbol{V}}^{k-1} - \boldsymbol{V}^{k-1})(l) \\ &\quad = \textstyle\sum_{s=0}^{k-1} \left\{ \bar{\boldsymbol{T}}^s h[\bar{r} - r](l) + \bar{\boldsymbol{T}}^s \left[(\bar{\boldsymbol{T}} - \boldsymbol{T})\boldsymbol{V}^{k-s-1}) \right] (l) \right\} + \bar{\boldsymbol{T}}^k(\bar{\boldsymbol{V}}^0 - \boldsymbol{V}^0)(l), \end{aligned} \tag{9.3.10}$$

where the last step followed by iteration. First note that the last term in the right hand side of (9.3.10) is equal to 0 as $\bar{\boldsymbol{V}}^0(\cdot) = \boldsymbol{V}^0(\cdot) = 0$. Furthermore, by (9.2.3) and (9.2.4), we can also write

$$\begin{aligned} &(\bar{\boldsymbol{T}} - \boldsymbol{T})\boldsymbol{V}^s(i) \\ &\quad = \sum_{j \neq i} h[\bar{\boldsymbol{q}}(i,j) - \boldsymbol{q}(i,j)]\boldsymbol{V}^s(j) - \sum_{j \neq i} h[\bar{\boldsymbol{q}}(i,j) - q(i,j)]\boldsymbol{V}^s(i) \\ &\quad = \sum_{j \neq i} h[\bar{\boldsymbol{q}}(i,j) - \boldsymbol{q}(i,j)][\boldsymbol{V}^s(j) - \boldsymbol{V}^s(i)]. \end{aligned} \tag{9.3.11}$$

By substituting (9.3.11) in (9.3.10) and noting that $\bar{\boldsymbol{T}}^s$ is a monotone operator for all s (i.e. $\bar{\boldsymbol{T}}^s f \leq \bar{\boldsymbol{T}}^s f$ if $f \leq g$ componentwise), from (9.3.10) and by using (9.3.7) we obtain:

$$(\bar{\boldsymbol{V}}^k - \boldsymbol{V}^k)(l) = \sum_{s=0}^{k-1} \bar{\boldsymbol{T}}^s \left\{[\bar{r}-r] + (\bar{\boldsymbol{T}}-\boldsymbol{T})\boldsymbol{V}^{k-s-1}\right\}(l) \geq 0. \tag{9.3.12}$$

The proof is completed by applying (9.3.3). □

Remark 9.3.3 (Essential difference to the stochastic comparison method) *As shown in section 9.2.2, with the stochastic comparison method or related sample path approach, as intensively studied in the literature, one requires that the one-change transition structure or rather the transition rate matrices $\boldsymbol{Q}$ and $\bar{\boldsymbol{Q}}$ are stochastically ordered as $\bar{\boldsymbol{Q}} \geq (\leq)\boldsymbol{Q}$ in some appropriate ordering sense. This would essentially imply condition (9.3.7) without the reward term $[\bar{r}-r]$.*

Such a strict ordering does seem quite natural in 'standard' type one-dimensional systems. However, for multi-dimensional queueing structures a strict ordering of purely the transition structure will be less natural and may not be satisfied. More precisely, the necessary ordering for the specific performance measure of interest might not be covered by the transition itself, as shown in section 9.2.3. By result 9.3.2, however, an ordering for that measure might still be provable by using the extra reward term $[\bar{r}-r]$ in condition (9.3.7). For the tandem example of section 9.2.3 this will be shown in section 9.5.

Remark 9.3.4 (Bias-terms) *The essential step to apply result 9.3.2 is to verify condition (9.3.7). This in turn will generally require to bound the so-called bias terms $\boldsymbol{V}^k(j) - \boldsymbol{V}^k(i)$ from below (or above) by 0. This bounding can be quite technical. But in general it can be performed in an inductive manner by exploiting the recursive relation (9.3.5). As these bias-terms will also play a crucial role to obtain error bounds, a more detailed discussion on these bias terms can be found in the next section.*

9.3.3 Error bound Result

Reconsider the setting of section 9.3.2 with

$$\begin{cases} \text{an original Markov reward chain } (\boldsymbol{S}, \boldsymbol{q}, r), \\ \text{an approximate Markov reward chain } (\bar{\boldsymbol{S}}, \bar{\boldsymbol{q}}, \bar{r}), \end{cases}$$

where both are assumed to be uniformizable with some constant H and where $\boldsymbol{S} \subseteq \boldsymbol{S}$. Let π and $\bar{\pi}$ denote their steady-state distributions. The following theorem can be given in various versions. The present form, however, is most practical in the natural situation that the steady-state distribution of one of the two models, typically the modified one, is known as easily computable. For convenience, we write

$$\bar{\pi} f = \sum_i \bar{\pi}(i) f(i)$$

Result 9.3.5 (Error bound) *Suppose that for some nonnegative function $\gamma(\cdot)$ at $\bar{\boldsymbol{S}}$, all $i \in \bar{\boldsymbol{S}}$ and $k \geq 0$:*

$$\left|[\bar{r}-r](i)+\sum_j[\bar{\boldsymbol{q}}(i,j)-\boldsymbol{q}(i,j)][\boldsymbol{V}^k(j)-\boldsymbol{V}^k(i)]\right| \leq \gamma(i). \tag{9.3.13}$$

Then

$$\left|\bar{\boldsymbol{G}}-\boldsymbol{G}\right| \leq \sum_i \bar{\pi}(i)\gamma(i)=\bar{\pi}\gamma. \tag{9.3.14}$$

Proof. Recall the derivation (9.3.10) for fixed $l \in \bar{\boldsymbol{S}}$ with the last term in the right hand side vanished as it is equal to 0. (as in the left hand side of (9.3.12)). Then by multiplication by $\pi(l)$ and summing over all l, we obtain

$$\begin{aligned}(\bar{\pi}\bar{\boldsymbol{V}}^k-\bar{\pi}\boldsymbol{V}^k) &= \sum_l \bar{\pi}(l)[(\bar{\boldsymbol{V}}^k-\boldsymbol{V}^k)(l)]\\ &= \sum_l \bar{\pi}(l)\sum_{s=0}^{k-1}\left\{h[\bar{r}-r](l)+[(\bar{\boldsymbol{T}}-\boldsymbol{T})\boldsymbol{V}^{k-s-1}(l)]\right\}\end{aligned} \tag{9.3.15}$$

Now note that since $\bar{\pi}$ (as steady state distribution) is invariable under $\bar{\boldsymbol{T}}$. For any function g:

$$\begin{aligned}\bar{\pi}(\boldsymbol{T}g) &= \sum_l \bar{\pi}(l)\sum_j \bar{\boldsymbol{P}}(l.j)g(j)\\ &= \sum_j\left[\sum_l \bar{\pi}(l)\bar{\boldsymbol{P}}(l,j)\right]g(j)=\sum_j \bar{\pi}(j)g(j)=\bar{\pi}g\end{aligned}$$

so that for any s:

$$\bar{\pi}\bar{\boldsymbol{T}}^s g=\bar{\pi}\bar{\boldsymbol{T}}(\bar{\boldsymbol{T}}^{s-1}g)=\bar{\pi}(\bar{\boldsymbol{T}}^{s-1}g)=\ldots=\bar{\pi}g \qquad (s>0) \tag{9.3.16}$$

As a consequence, by also taking absolute values

$$\begin{aligned}\left|\bar{\pi}\bar{\boldsymbol{V}}^k-\bar{\pi}\boldsymbol{V}^k\right| &= \left|\textstyle\sum_{s=0}^{k-1}\bar{\pi}\left\{h[\bar{r}-r]+[\bar{\boldsymbol{T}}-\boldsymbol{T}]\boldsymbol{V}^{k-s-1}\right\}\right|\\ &\leq k\textstyle\sum_i \bar{\pi}(i)\left|h[\bar{r}-r](i)+[\bar{\boldsymbol{T}}-\boldsymbol{T}]\boldsymbol{V}^{k-s-1}(i)\right|\end{aligned} \tag{9.3.17}$$

Substitution of (9.3.17) into (9.3.15) and using condition (9.3.13) thus gives

$$\left|\bar{\pi}\bar{\boldsymbol{V}}^k-\bar{\pi}\boldsymbol{V}^k\right| \leq kh\sum_i \bar{\pi}(i)\gamma(i)=kh[\bar{\pi}\gamma] \tag{9.3.18}$$

By recaling the steady-state convergence (9.3.3) to be independent of the initial state, the proof is thus completed by

$$\begin{cases} \dfrac{H}{k}\sum_i \bar{\pi}(i)\bar{\boldsymbol{V}}^k(i) \to \bar{\boldsymbol{G}}\ (k\to\infty) \\ \dfrac{H}{k}\sum_i \bar{\pi}(i)\boldsymbol{V}^k(i) \to \boldsymbol{G}\ (k\to\infty) \end{cases}$$

□

Remark 9.3.6 (Application of result 9.3.5) *Note that result 9.3.5 may lead to small error bounds in either of two ways:*

- *When either the difference between the transition rates $\boldsymbol{q}$ and $\bar{\boldsymbol{q}}$ is small, uniformly in all states, say $\|\bar{\boldsymbol{Q}} - \boldsymbol{Q}\| \le \varepsilon$ for some small ε. Here one may typically think of small **perturbations or inaccuracies** in system parameters such as an arrival rate λ. (Examples of this form for $M|M|c$-systems can be found in [23]).*
- *When the transition rates $\boldsymbol{q}$ and $\bar{\boldsymbol{q}}$ may differ quite strongly in specific states i, but where the likelihood $\bar{\pi}(i)$ of being in such states is rather small. Here, one could typically think of a system **modification or truncation**. In section 9.4 a situation (an error bound for a modification) is illustrated for the instructive example from section 9.1.1.2. Below (in section 9.3.4) the specific case of a truncation will be made more explicit. In section 9.6 a truncation error bound will be obtained for finite Jackson networks.*

Remark 9.3.7 (Bounded bias-terms) *A crucial step to apply results 9.3.5 (as well as result 9.3.2) is to bound the difference terms (in stochastic dynamic programming also known as relative gain or bias-terms) of the form:*

$$[\boldsymbol{V}^k(j) - \boldsymbol{V}^k(i)] \tag{9.3.19}$$

Clearly, $\boldsymbol{V}_t$ will generally grow linearly in t and thus be unbounded. However, the difference term (9.3.19) for fixed i, j will generally be bounded regardless of t.

More precisely, when r is bounded, say $\|r\| < M$, by simple Markov reward arguments (cf. [23]) one proves

$$|\boldsymbol{V}^k(j) - \boldsymbol{V}^k(i)| \le 2M\, min[R_{ij}, R_{ji}] \tag{9.3.20}$$

where R_{ij} is the expected number of steps (mean first passage time; see [42] to reach state j out of i. A similar though more technical result in terms of these times can be given also for unbounded rewards (cf. [12]). Most essentially, however, closed-form expressions or simple bounds for mean first passage times seem to be limited to simple one-dimensional random walks (cf. [42]). In the next sections, however, we will demonstrate how bounds for bias terms can be established in an analytic manner by inductive Markov reward arguments or more precisely by employing the dynamic reward relation (9.3.5).

Remark 9.3.8 (Bounds for bias-terms) *For result 9.3.2 only lower (≥ 0) or upper estimates (≤ 0) by zero for the bias-terms are needed. However, as might turn out* *in more complicated situations, also absolute bounds might be required to prove these 0-estimates. This will appear and be illustrated for the finite tandem example in section 9.5.*

For result 9.3.5 these bounds will be used in the absolute errors $\gamma(i)$ (conversely also the 0-lower or upper estimates might still be required in the inductive proofs to compensate for the extra reward terms).

Conveniently, by condition (9.3.7) in result 9.3.2 and (9.3.13) in result 9.3.5, bounds for (9.3.19) are only required for 'neighboring' states for which

$$|\bar{\boldsymbol{q}}(i,j) - \boldsymbol{q}(i,j)| > 0. \tag{9.3.21}$$

For example in a birth-death queueing system, only for states $j = i-1$ or $j = i+1$.

For standard type queueing networks (that is, without batch movements) only for states of the form:

$$\begin{cases} i = \boldsymbol{n} \\ j = \boldsymbol{n} + \boldsymbol{e}_p - \boldsymbol{e}_q \text{ with } \boldsymbol{n} = (n_1, \ldots, n_J) \end{cases} \tag{9.3.22}$$

representing the population numbers n_s at each stations to indicate that only one job has moved form one station p to another station q. This natural queueing network 'property' of one job-shift at a time can often be exploited to provide analytic bounds for (9.3.19), as will be illustrated in sections 9.4, 9.5 and 9.6.

Remark 9.3.9 ($\bar{\boldsymbol{S}} = \boldsymbol{S}$) *Note that the role of the original and modified system can nót be interchanged when the state spaces are not equal, i.e. $\bar{\boldsymbol{S}} \subsetneq \boldsymbol{S}$. However, for $\bar{\boldsymbol{S}} = \boldsymbol{S}$ and by using that*

$$\sum_i \pi(i)(\bar{\boldsymbol{T}}^s f)(i) \xrightarrow{C} \sum_j \bar{\pi}(j) f(j)$$

in combination with limiting arguments, in the proof of the theorem we can also replace the steady state distribution $\bar{\pi}$ by the steady state distribution π. Hence, in that case (9.3.14) can be read either with $\bar{\pi}$ or with $\bar{\pi}$ replaced by π. Alternatively, by reversing the roles, when $\bar{\boldsymbol{S}} = \boldsymbol{S}$, we may also read result 9.3.5 with (9.3.13) replaced by

$$\left| [\bar{r} - r](i) + \sum [\bar{\boldsymbol{q}}(i,j) - \boldsymbol{q}(i,j)][\bar{\boldsymbol{V}}^k(j) - \bar{\boldsymbol{V}}^k(i)] \right| \leq \gamma(i) \tag{9.3.23}$$

while keeping (9.3.14) as it stands. This alternative condition provides more flexibility to use the more appropriate system, either the original or modified one, in order to establish bounds for the bias-terms. (This will appear to be convenient later on for an application in section 9.4.3)

9.3.4 Truncation Error Bound

As a special case, as of interest by itself, in this section consider the situation of a truncation such as to jusitfy:

- a reduction of numerical effort, or conversely,
- the approximation of a finite model by a solvable infinite model.

To this end, let

$$\begin{cases} \bar{\boldsymbol{S}} \subseteq \boldsymbol{S} \\ \bar{r} = r \text{ and} \\ \bar{\boldsymbol{q}}(i,j) = \begin{cases} 0 & (j \notin \bar{\boldsymbol{S}}) \\ \boldsymbol{q}(i,j) + \sum_{k \notin \bar{\boldsymbol{S}}} 1_{\{t[i,k]=j\}} \boldsymbol{q}(i,k) & (j \in \bar{\boldsymbol{S}}) \end{cases} \end{cases} \tag{9.3.24}$$

In words that is, a transition from i to k by which $\bar{\boldsymbol{S}}$ would be left is transformed into a transition into a special truncation state $\boldsymbol{t}[i,k] \in \bar{\boldsymbol{S}}$. The notation and quantities as defined before are adopted with an upper bar for the truncated model.

The following truncation result 9.3.10 now follows almost directly from result 9.3.5. This result roughly states that the effect of a state space truncation can be expressed by a steady state weight of the one step effect of the truncation on the bias-terms. If the probability mass for states in which the truncation does have a direct effect is small, also the error bound can be expected to be small.

Result 9.3.10 *Suppose that for some function γ at $\bar{\boldsymbol{S}}$, all $k \geq 0$ and any state $i \in \bar{\boldsymbol{S}}$:*

$$\left| \sum_{j \notin \bar{\mathrm{S}}} \boldsymbol{q}(i,j) \left[\boldsymbol{V}^k(j) - \boldsymbol{V}^k(\boldsymbol{t}[i,j]) \right] \right| \leq \gamma(i)\,. \tag{9.3.25}$$

Then

$$\left| \bar{\boldsymbol{G}} - \boldsymbol{G} \right| \leq \sum_{i \in \bar{\boldsymbol{S}}} \bar{\pi}(i)\gamma(i) = \bar{\pi}\gamma. \tag{9.3.26}$$

Proof. For $i \in \bar{\boldsymbol{S}}$ we have:

$$[\bar{\boldsymbol{q}}(i,j) - \boldsymbol{q}(i,j)][\boldsymbol{V}^k(j) - \boldsymbol{V}^k(i)] =$$

$$\begin{cases} -\boldsymbol{q}(i,j)[\boldsymbol{V}^k(j) - \boldsymbol{V}^k(i)] & ,(j \notin \bar{\boldsymbol{S}})\,. \\ \left[\sum_{\{k \notin \bar{\boldsymbol{S}} \,|\, t[i,k]=j\}} \boldsymbol{q}(i,k) \right] [\boldsymbol{V}^k(j) - \boldsymbol{V}^k(i)] & ,(j \in \bar{\boldsymbol{S}})\,. \end{cases} \tag{9.3.27}$$

with $\bar{r} = r$, condition (9.3.13) thus reduces to

$$\sum_j [\bar{\boldsymbol{q}}(i,j) - \boldsymbol{q}(i,j)][\boldsymbol{V}^k(j) - \boldsymbol{V}^k(i)] = \sum_{l \in \boldsymbol{S}/\bar{\boldsymbol{S}}} \boldsymbol{q}(i,l)[\boldsymbol{V}^k(t[i,l]) - \boldsymbol{V}^k(l)]$$

Result 9.3.5 completes the proof. □

Remark 9.3.11 (≥ 0 or ≤ 0) *Clearly, depending on the reward rate (performance measure of interest), as in result 9.3.5 also in (9.3.25) an inequality sign* ≥ 0 *(or* ≤ 0*) can be included, so as to conclude (or similarly with* $\boldsymbol{G}$ *and* $\bar{\boldsymbol{G}}$ *interchanged)*

$$0 \leq \boldsymbol{G} - \bar{\boldsymbol{G}} \leq \sum_{i \in \bar{\boldsymbol{S}}} \bar{\pi}(i)\gamma(i), \tag{9.3.28}$$

Remark 9.3.12 (Infinite expansion) *As a special case, the* $\boldsymbol{q}$*-model might also represent an infinite model. Particularly, thinking of a queueing network application, the* $\boldsymbol{q}$*-model may correspond to an infinite product form approximation for a finite non-product form system. This will be used in section 9.6.*

Remark 9.3.13 (Computational error bound) *Note that the error bound in (9.3.26) necessarily requires the steady state distribution* $\bar{\pi}$ *for the system with the smallest state space* $\boldsymbol{S} \subseteq \boldsymbol{S}$ *(as essentially used in the proof of result 9.3.5). This distribution may typically be thought of as only obtainable by numerical computation.*

Remark 9.3.14 (Analytic error bound) *The computational error bound in (9.3.26), in turn, might be estimated form above by an analytic expression of the form*

$$\sum_i \bar{\pi}(i)\gamma(i) \leq \sum_i \bar{\bar{\pi}}(i)\gamma(i) \tag{9.3.29}$$

by using an analytic approximation $\bar{\bar{\pi}}$ *and the stochastic comparison result 9.2.1. In this form, the stochastic comparison and Markov reward approach might become mutually beneficial.*

Particularly, for queueing network applications an analytic error bound of the form (9.3.29) can be thought of as by a product form modification or infinite expansion. In section 9.6.3 this will be established for finite Jackson networks.

9.3.5 Comparison of MRA and SC

Clearly both approaches of the Markov reward approach (MRA), as presented in this section, and of stochastic comparison (SC), as briefly presented in section 9.2.1 (which in turn is directly related to sample path comparison) and which has been studied intensively in the literature, have a common starting point and objective of:

> *Comparing the performance of two related stochastic systems*

As such, a comparison of the two approaches is in place on a number of aspects each of which might lead to an advantage (or preference) or just the opposite: a disadvantage (or limitation) for either of the two approaches.

1. **Objective: ordering and error bound.** First of all, the MRA might lead to both an ordering and a quantification by means of an error bound for the discrepancy of the two systems. SC only establishes ordering results (clearly, in specific applications, as shown and discussed for an $M|G|1$-application in [20], ordering results in combination with analytic expressions for bounding models might indirectly also lead to an error bound).

   These error bounds in turn might typically be thought of as being small, where one system is a modification of the other, for either of two reasons:

   - when the modification itself is small
   - when the likelihood for the modification to take place is small

   In principle, the error bounds do not (need to) rely upon ordering results. Nevertheless, in applications usually also ordering results are included as side results.

   (As an example, in [22], for a non-exponential extension of the MRA, error bounds are obtained for $GI|G|c$-systems with different service hazard rates, despite the fact that these hazard rates themselves are not ordered, as illustrated in figure 9.4.)

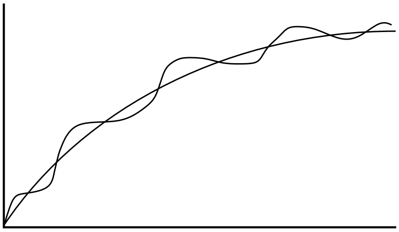


Fig. 9.4: Nón-ordered hazard rates for two $M|G|c$ systems.

   In fact, the opposite applied here, by the error bound also an ordering result for the systems could be concluded.)

2. **Exponential case: essential technical difference.** For the exponential case, in essence, as shown in section 9.2.1, by SC one shows that the one-step transition matrices $\boldsymbol{P}$ and $\bar{\boldsymbol{P}}$ (or rather the generators $\boldsymbol{Q}$ and $\bar{\boldsymbol{Q}}$) are ordered. By the MRA, in contrast, an ordering (as well as quantification) is required (see relation (9.3.7) or (9.3.13)) for the combination of the one-step transition matrix (or generator) and the one-step reward function r as by

$$[r+\boldsymbol{P}\boldsymbol{V}^k] \quad \text{and} \quad [\bar{r}+\bar{\boldsymbol{P}}\bar{\boldsymbol{V}}^k]$$

   This technical difference by itself will imply a number of differences in favour of one of the two approaches as will be mentioned below.

3. **’Stronger’ comparison result (1)** SC thus leads to ordering results that hold for all possible reward functions (and related performance measures) from a, generally wide, monotonicity class $\mathbb{M}$ (as in section 9.2.2, whereas the MRA only deals with one specific reward rate (performance measure). In this sense, SC can be regarded as being stronger.

   However, as a price to pay, also the implicit conditions for the system to satisfy a stochastic comparison ordering will generally be stronger!

   Particularly, stochastic comparison results have been reported widely for multi-server type queues, which can be seen as one-dimensional systems. For more complicated, say multi-dimensional, queueing systems, however, such as queueing networks, stochastic comparison results are far more limited but can still be obtained as in [63] and the chapter by Szekli (also see remark 9.6.5).

   More concrete, as shown in section 9.2.1, to establish an ordering result for a specific performance measure, SC might fail, while the MRA, as will be shown in section 9.5, might still work, based upon the specific reward function r of interest. In this respect, also the MRA might be referred to as ’stronger’.

4. **’Stronger’ comparison result (2).** Somewhat relatedly, SC or the related sample path comparison approach does in fact lead to ordering results that even apply at sample path basis (with probability 1). The MRA in principle only provides a comparison at expected steady state basis (see section 9.7 for an extension to the transient case). Again, as such SC can be regarded as ’stronger’.

   However, as can be shown easily by ’counterintuitive’ examples (such as in sections 9.4.1 and section 9.6.2, ordering results might fail at sample path basis but still be expected and be proven by the MRA at expected steady state basis. This will be shown in sections 9.4 and 9.6.

   (Another illustration can be found in [64] which shows an ordering result for an availability measure at expected steady state basis while the underlying systems are not, neither in strong nor weak sense, stochastically ordered).

   In this respect, also the MRA might again be referred to as ’stronger’.

5. **’Proofs’.** The proofs for stochastic comparison or ordering results, as often given by sample path comparison and weak coupling arguments, are generally most elegant. The technical verifications for the MRA, in contrast, in particular for the estimation of the bias-terms (9.3.19), are generally more complicated and technical. On the other hand, these verifications are often also more structured. As a consequence and in line with 3, in more complex situations the MRA might be favourable if not necessary.

6. **(Non)exponentiality.** The exponentiality requirement of the MRA is a strong limitation for practical applications. SC or rather the approach by sample path comparison, in contrast, in general works, if it applies, for queueing systems with arbitrary service and arrival distributions.

   Various non-exponential extensions of the MRA for specific applications have meanwhile been developed in the literature, by using phase-type distributions (e.g. see [22]). Nevertheless, the technicalities become far more complicated. In this respect, SC remains to be highly favourable.

7. **Combination of both.** In fact, it might also be beneficial to combine the MRA and SC so as to eventually obtain a simple analytic expression for an error bound. This will be illustrated in section 9.6.3 for the truncation and expansion of finite Jackson networks.

Table 9.3: Overview.

A somewhat 'imprecise' and merely 'global' overview of these reflections is listed in table 9.3. The letters **A** and **D** indicate whether this aspect should generally be seen as an advantage or a disadvantage.

Markov Reward Approach		**Stochastic Comparison**	
Error Bound Results	(**A**)	-	
Comparison results		Comparison results	
More complex systems	(**A**)	Strong system conditions	(D)
Queueing networks	(**A**)	'Simple' queueing systems	(D)
Might still work	(**A**)	Might not apply	(D)
Only one measure	(D)	Class of measures	(**A**)
Only at expectation basis	(D)	Sample path results	(**A**)
Exponential requirements	(D)	No exponentiality required	(**A**)
Technical analytic proofs	(D)	Elegant sample path proofs	(**A**)
Comparison as extra	(**A**)		
Combination			

B: Applications

9.4 Application 1: Instructive Breakdown Example

This section aims to illustrate the Markov reward approach and the results from section 9.3 in an instructive manner. More precisely, it will illustrate:

1. How the bias-terms (9.3.19) for different performance measures can be estimated analytically.
2. How the conditions (9.3.7) and (9.3.13) or (9.3.25) can be verified.
3. The type of results that can be obtained.

To this end, reconsider the instructive breakdown example from section 9.1.1.2. As argued in remarks 9.3.4 and 9.3.7, a crucial step to apply the Markov reward approach is to estimate the bias-terms (9.3.19) for the measure of interest. In section 9.4.1, therefore, we first show how this can be achieved in an analytic manner using the reward relation (9.3.5).

Next, in sections 9.4.2 and 9.4.3, a comparison and two error bound applications are given to show how the results from section 9.3 can be used, once bounds for these bias-terms have been established.

9.4.1 Analytic bounds for the bias-terms

Typical performance measures of interest are:

- a throughput
- a mean queue length
- or directly related measures as a loss probability or a delay

In this section it will be shown how comparison and error bound results as in section 9.3 can be investigated for different measures in an analytic and more or less unified manner . The common first step is the derivation of a recursive relation for the bias-terms by using the reward relation (9.3.5) and comparing the possible transitions in two neighboring states. More precisely, with (n,θ) as in section 9.1.1.2, with $n \leq N$ the number of jobs and $\theta = 1,0$ the status of the server being up ($\theta = 1$) or down ($\theta = 0$), we aim to obtain an analytic expression for

$$[V^k(n+1,\theta) - V^k(n,\theta)]$$

This expression can be obtained in an analytic manner by subtracting the reward expression (9.3.5) in the state (n,θ) from the reward expression (9.3.5) in state

$(n+1,\theta)$. However, the derivation of these bias-term expressions are generally understood more easily in a stochastic manner by comparing each of the transitions, which can take place in either of the two states, pairwise. This will be illustrated in more detail below for the breakdown example from section 9.1.1.2. Let $r(\cdot)$ be the reward rate for the performance measure of interest (such as by (9.2.6)) and consider a fixed k. Let

$$H = [\lambda + \mu + \gamma_0 + \gamma_1]$$

Expression for $[\boldsymbol{V}(n+1,\theta) - \boldsymbol{V}(n,\theta)]$**.**

By substituting the transition rates $\boldsymbol{q}((n,\theta),(n,\theta)')$ and with $h = H^{-1}$, the uniformized transition matrix $\boldsymbol{P}$ as by (9.2.3) becomes:

$$\boldsymbol{P}((n,\theta),(n,\theta)') =$$
$$\begin{cases} & \underline{(n,\theta)'} \\ h\lambda 1_{(n<N)} & ,(n+1,\theta) \\ h\lambda 1_{(n>0)} 1_{(\theta=1)} & ,(n-1,1) \\ h\gamma_1 1_{(\theta=1)} + h\gamma_0 1_{(\theta=0)} & ,(n,[\theta+1](mod 2)) \\ \left[1 - h\lambda 1_{(n<N)} - h\mu 1_{(n>0)} 1_{(\theta=1)} - h\gamma_\theta\right] & ,(n,\theta)(n \leq N, \theta = 0,1) \end{cases} \tag{9.4.1}$$

By the reward relation (9.3.5) in state $i = (n,\theta)$, for $k+1$, we then obtain:

$$\begin{aligned} \boldsymbol{V}^{k+1}(n,\theta) = {} & hr(n,\theta) \\ & + h\lambda 1_{(n<N)} \boldsymbol{V}^k(n+1,\theta) \\ & + h\mu 1_{(n>0)} 1_{(\theta=1)} \boldsymbol{V}^k(n-1,1) \\ & + h\gamma_1 1_{(\theta=1)} \boldsymbol{V}^k(n,0) + h\gamma_0 1_{(\theta=0)} \boldsymbol{V}^k(n,1) \\ & + [1 - h\lambda 1_{(n<N)} - h\mu 1_{(n>0)} 1_{(\theta=1)} - h\gamma_\theta] \boldsymbol{V}^k(n,\theta) \end{aligned} \tag{9.4.2}$$

Similarly, in state $(n+1,\theta)$ with $n+1 \leq N$ and for $k+1$, we obtain:

$$\begin{aligned} \boldsymbol{V}^{k+1}(n+1,\theta) = {} & hr(n+1,\theta) \\ & + h\lambda 1_{(n+1<N)} \boldsymbol{V}^k(n+2,\theta) \\ & + h\mu 1_{(\theta=1)} \boldsymbol{V}^k(n,1) \\ & + h\gamma_1 1_{(\theta=1)} \boldsymbol{V}^k(n+1,0) + h\gamma_0 1_{(\theta=0)} \boldsymbol{V}^k(n+1,1) \\ & + [1 - h\lambda 1_{(n+1<N)} - h\mu 1_{(n>0)} 1_{(\theta=1)} - h\gamma_\theta] \\ & \quad \boldsymbol{V}^k(n+1,\theta) \end{aligned} \tag{9.4.3}$$

Now, in order to subtract (9.4.2) from (9.4.3) in states with $n+1 \leq N$, hence $n < N$, and to compare the transitions in a pairwise manner, in the right hand side of (9.4.2), rewrite:

$$\begin{aligned} & h\lambda 1_{(n<N)} \boldsymbol{V}^k(n+1,\theta) \\ & \quad = h\lambda 1_{(n+1<N)} \boldsymbol{V}^k(n+1,\theta) \\ & \quad + h\lambda 1_{(n+1=N)} \boldsymbol{V}^{k+1}(n+1,\theta) \end{aligned}$$

as well as (artificially add and subtract) include the extra term (which is equal to 0):

$$h\mu 1_{(n=0)} 1_{(\theta=1)} \boldsymbol{V}^k(n,1) - h\mu 1_{(n=0)} 1_{(\theta=1)} \boldsymbol{V}^k(n,1)$$

This 0-term is included as if in state (n,θ) with $n=0$ there is also a 'dummy transition' with probability $h\mu 1_{(\theta=1)}$, by which the state remains unchanged, as there is a transition with this probability in (9.4.3) for state $(n+1,\theta)$. Conversely, for a similar reason, in state $(n+1,\theta)$ with $n+1=N$, in the right hand side of (9.4.3), we (artificially add and subtract) include the extra term (which is equal to 0):

$$h\lambda 1_{(n+1=N)} \boldsymbol{V}^{k+1}(n+1,\theta) - h\lambda 1_{(n+1=N)} \boldsymbol{V}^{k+1}(n+1,\theta)$$

Then, after these substitutions have been made and by subtracting (9.4.2) from (9.4.3), for any state (n,θ) with $n+1 \leq N$ and $\theta = 0,1$, we find

$$\begin{aligned} & \left[\boldsymbol{V}^{k+1}(n+1,\theta) - \boldsymbol{V}^{k+1}(n,\theta)\right] \\ & \quad = h\left[r(n+1,\theta) - r(n,\theta)\right] && (4.4.1) \\ & \quad + h\lambda 1_{(n+1<N)} \left[\boldsymbol{V}^k(n+2,\theta) - \boldsymbol{V}^k(n+1,\theta)\right] && (4.4.2) \\ & \quad + h\lambda 1_{(n+1=N)} \left[\boldsymbol{V}^k(N,\theta) - \boldsymbol{V}^k(N,\theta)\right] && (4.4.3) \\ & \quad + h\mu 1_{(n>0)} 1_{(\theta=1)} \left[\boldsymbol{V}^k(n,\theta) - \boldsymbol{V}^k(n-1,\theta)\right] && (4.4.4) \qquad (9.4.4) \\ & \quad + h\mu 1_{(n=0)} 1_{(\theta=1)} \left[\boldsymbol{V}^k(0,1) - \boldsymbol{V}^k(0,1)\right] && (4.4.5) \\ & \quad + h\gamma_1 1_{(\theta=1)} \left[\boldsymbol{V}^k(n+1,0) - \boldsymbol{V}^k(n,0)\right] && (4.4.6) \\ & \quad + h\gamma_0 1_{(\theta=0)} \left[\boldsymbol{V}^k(n+1,1) - \boldsymbol{V}^k(n,1)\right] && (4.4.7) \\ & \quad + \left[1 - h\lambda - h\mu 1_{(\theta=1)} - h\gamma_\theta\right] \left[\boldsymbol{V}^k(n+1,\theta) - \boldsymbol{V}^k(n,\theta)\right] && (4.4.8) \end{aligned}$$

Here indeed, the terms (4.4.3) and (4.4.5) in the right hand side of (9.4.4) are equal to 0 but left in for clarity of the derivation as well as an argument that will follow below in the proof of lemma 9.4.1. The relation (9.4.4) can now be used to obtain an analytic lower and upper bound for this bias-terms.

Lemma 9.4.1 (Throughput) *Let $r(n) = \mu 1_{(n>0)} 1_{(\theta=1)}$. Then for all $k \geq 0$, $n < N$ and $\theta = 0,1$:*

$$0 \leq \Delta \boldsymbol{V}^k(n,\theta) = \left[\mathbf{V}^k(n+1,\theta) - \mathbf{V}^k(n,\theta)\right] \leq 1 \tag{9.4.5}$$

Proof. This will follow by induction in k. Clearly, (9.4.5) holds for $k = 0$. Assume that (9.4.5) is satisfied for all $k \leq l$. Then by (9.4.4) (keeping in the fourth term from the right hand side, which is equal to 0) and substituting $r(n) = \mu 1_{(n>0)} 1_{(\theta=1)}$ for $k = l + 1$ we obtain:

$$\begin{aligned}
&\Delta V^{l+1}(n,\theta) \\
&\quad = h\mu 1_{(n=0)} 1_{(\theta=1)} \\
&\quad + h\lambda 1_{(n+1<N)} \Delta V^l(n+1,\theta) \\
&\quad + h\mu 1_{(n>0)} 1_{(\theta=1)} \Delta V^l(n-1,1) \\
&\quad + h\mu 1_{(n=0)} 1_{(\theta=1)} \left[V^l(0,1) - V^l(0,1)\right] \\
&\quad + h\gamma_1 1_{(\theta=1)} \Delta V^l(n,0) h\gamma_0 1_{(\theta=0)} \Delta V^l(n,1) \\
&\quad + \left[1 - h\lambda - h\mu 1_{(\theta=1)} - h\gamma_\theta\right] \Delta V^l(n,\theta)
\end{aligned} \tag{9.4.6}$$

By substituting the induction hypothesis $\Delta V^l(n,\theta) \geq 0$ for all (n,θ), we can estimate the right hand side of (9.4.6) from below by 0 and directly conclude: $V^l(n,\theta) \geq 0$. To estimate the right hand side of (9.4.6) from above, now first observe that the first additional term $h\mu 1_{(n>0)} 1_{(\theta=1)}$ is equal to the (probability) coefficient of the fourth term which is equal to 0. Furthermore, note that all the coefficients represent transition probabilities that do not add up to more than 1. As a consequence, by substituting the induction hypothesis: $\Delta V^l(n,\theta) \leq 1$ for all (n,θ) and adding up all these coefficients, we can estimate the right hand side of (9.4.6) from above and conclude: $\Delta V^{l+1}(n,\theta) \leq 1$ for all (n,θ). We have thus proven (9.4.6) for $k = l + 1$. The induction completes the proof. □

For the comparison and truncation application in sections 9.4.2 and 9.4.3 the throughput is the measure of first interest. Lemma 9.4.1 can then be applied. For the error bound applications in section 9.4.3, it is also of particular interest to consider a mean queue length. In this case, it appears to be more convenient to reverse the roles of the original and modified system. That is, we investigate the bias-terms of the modified breakdown system as described under (v) in section 9.1.1.2, that is with arrivals also rejected when the system is down. This will be used in lemma 9.4.2 below.

Lemma 9.4.2 (Mean queue length) *Let $r(n,\theta) = n$. Then with arrival rejection when the system is down, for all $k \geq 0$, $n < N$ and $\theta = 0, 1$:*

$$0 \leq \Delta V^k(n,\theta) \leq \frac{[n+1]}{[\mu - \lambda]} \tag{9.4.7}$$

Proof. As for lemma 9.4.1 this will follow by induction in k. Clearly, (9.4.7) holds for $k = 0$ Assume that (9.4.7) holds for $k \leq l$. Then by (9.4.4) (with the addition of arrivals rejected when the system is down and again with the fifth term, which is equal to 0, kept in), and by substituting $r(n) = n$, we find:

$$\begin{aligned}
&\Delta V^{l+1}(n,\theta)\\
&\quad = h\\
&\quad + h\lambda 1_{(n+1<N)}\lambda 1_{(\theta=1)}\Delta V^{l}(n+1,\theta)\\
&\quad + h\mu 1_{(n>0)}1_{(\theta=1)}\Delta V^{l}(n-1,\theta)\\
&\quad + h\mu 1_{(n=0)}1_{(\theta=1)}\left[V^{l}(0,1)-V^{l}(0,1)\right]\\
&\quad + h\gamma_1 1_{(\theta=1)}\Delta V^{l}(n,0)h\gamma_0 1_{(\theta=0)}\Delta V^{l}(n,1)\\
&\quad + \left[1-h\lambda 1_{(n+1<N)}1_{(\theta=1)}-h\mu 1_{(\theta=1)}-h\gamma_1-h\gamma_0\right]\Delta V^{l}(n,\theta)
\end{aligned} \tag{9.4.8}$$

Again, by substituting the lower estimates 0 by the induction hypothesis (9.4.7) for $k=l$, by (9.4.8) one directly verifies $V^{l+1}(n,\theta)\geq 0$. To estimate the right hand side of (9.4.8) from above, substitute $\Delta V^{l}(n,\theta)\leq[n+1]C$ in order to prove that $\Delta V^{l+1}(n,\theta)\leq[n+1]C$. We then require that

$$\begin{aligned}
&h\left[1+\lambda 1_{(n+1<N)}1_{(\theta=1)}[n+2]C+h\mu 1_{(\theta=1)}nC+\gamma_\theta nC\right]\\
&\leq\left[1-\lambda 1_{(n+1<N)}1_{(\theta=1)}-h\mu 1_{(\theta=1)}-h\gamma_\theta\right][n+1]C\\
&\leq[n+1]C
\end{aligned}$$

This in turn, is satisfied by

$$1+\lambda C-\mu C\leq 0$$

Hence, by choosing $C=1/(\mu-\lambda)$ we have also proven that $\Delta V^{l+1}(n,\theta)\leq[n+1]/[\mu-\lambda]$. The induction completes the proof. □

In fact, lemma 9.4.2 and its proof can directly be reread in the more general form of:

Lemma 9.4.3 (General bounded case) *Let $r(n)$ be such that $0\leq[r(n+1,\theta)-r(n,\theta)]\leq\boldsymbol{R}$ for some constant $\boldsymbol{R}$. Then, for the system as in lemma 9.4.2, all $k\geq 0$, $n<N$ and $\theta=0,1$:*

$$0\leq\Delta V^{k}(n,\theta)\leq\boldsymbol{R}\frac{[n+1]}{[\mu-\lambda]}$$

Remark 9.4.4 (Other performance measures) *For specific applications bounds for the corresponding bias-terms have also been established in the literature for other measures such as for tail probabilities of tandem queues in [21] and for the availability (number of up components) of performability models in [64].*

9.4.2 Comparison Result

Consider the loss probability $\boldsymbol{B}=\pi(n,1)+\pi(n,0)$. As there is no analytic solution for B, as argued in section 9.1.1.2, one might suggest to use

$$\boldsymbol{B}_L \le \boldsymbol{B} \le \boldsymbol{B}_U \tag{9.4.9}$$

with

- $\boldsymbol{B}_L$: the loss probability for the system without breakdowns (or equivalently the system that continues to work also when the system is down) (called lower bound system).
- $\boldsymbol{B}_U$: the loss probability for the system in which arrivals are rejected when the system is down (called upper bound system) as by the product form expression (9.1.2).

As indicated by the numerical results in table 2.1 (Chapter 1) and as can be expected, these bounds can even be reasonably accurate for τ reasonably small, with

τ : the fraction of time that the system is down

Intuitively, the inequalities (9.4.9) seem trivial. Nevertheless, one has to be careful as shown by the following example. A formal comparison proof for (9.4.9) will thus be of interest. This will be established by result 9.4.6 below.

Example 9.4.5 *Let $N = 2$ and consider a processor sharing service discipline. Hence, with 2 jobs present each receives a service capacity $\frac{1}{2}$. (Note that this doesn't effect the total service rate μ as the service times are exponential).*

Let a realization of inter-arrival times and service requirements be given by:

Job	*1*	*2*	*3*	*4*	*5*
Arrival time	*3*	*7*	*11*	*11.5*	*22*
Service time	*5*	*3*	*1*	*2*	*6*

while the up and down times are:

Up	*Down*	*Up*	*Down*
6	*2*	*7*	*3*

The realizations in the original model (OM) and the upper bound model (UM) under which arrivals were rejected when the system is down are depicted in figure 9.5 by:

◎ *: arrival*
⊗ *: rejection*
□ *: completion*
D_i *: departure of ith accepted job*

We observe that the second arrival is accepted in the original model but rejected in the upper-bound model. This, however, leads to rejections later on for the original model at times 11 and 11.5 while in the upper-bound model the jobs are accepted as

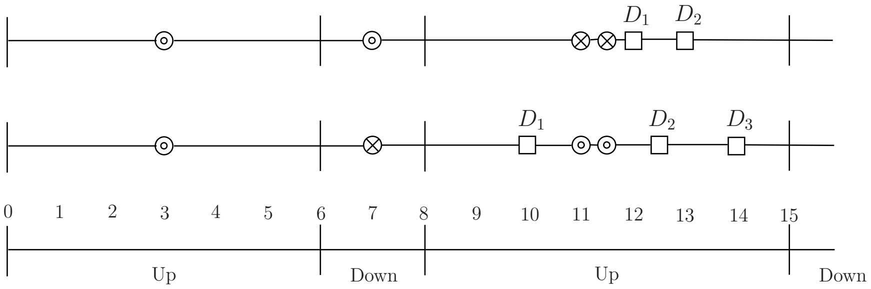


Fig. 9.5: Network Example

the system has become empty. During the time interval 0-15 the upper-bound model thus appears to have a better performance (smaller number of losses and larger number of completions) in contrast with the intuition that it will perform less.

Result 9.4.6

$$\boldsymbol{B}_L \leq \boldsymbol{B} \leq \boldsymbol{B}_U \tag{9.4.10}$$

Proof. By virtue of the relationship for the throughput $\boldsymbol{F} = \lambda(1-\boldsymbol{B})$, and similarly for the lower and upper bound system, it suffices to prove that $\boldsymbol{F}_L \geq \boldsymbol{F} \geq \boldsymbol{F}_U$. Let us restrict to the upper bound system. We will apply result 9.3.2 and lemma 9.4.1.

Let $\bar{\boldsymbol{q}}$ correspond to the upper bound system and $\boldsymbol{q}$ to the original one. To verify condition (9.3.7), first note that $r = \bar{r}$. Furthermore, note that

$$\begin{aligned}\sum_{(n,\theta)'} \left[\bar{\boldsymbol{q}}\left((n,\theta),(n,\theta)'\right) - \boldsymbol{q}\left((n,\theta),(n,\theta)'\right)\right] \left[\boldsymbol{V}^k\left((n,\theta)'\right) - \boldsymbol{V}^k\left((n,\theta)\right)\right] \\ = -\lambda 1_{(n<N)} 1_{(\theta=0)} \left[\boldsymbol{V}^k(n+1,0) - \boldsymbol{V}^k(n,0)\right]\end{aligned} \tag{9.4.11}$$

By lemma 4.1, the right hand side of (9.4.11) can be estimated from above by 0. By result 9.3.2 (with the signs reversed), we thus conclude $\boldsymbol{F}_U \leq \boldsymbol{F}$. (The proof for the lower bound $\boldsymbol{B}_L$ follows similarly if we let $\bar{\boldsymbol{q}}$ correspond to the system without breakdowns or equivalently the system that always continues to work, also when the system is down.) □

9.4.3 Error Bounds

Despite the fact the lower and upper bound systems have explicit product form expressions, let us also investigate whether we can quantify their inaccuracy by analytic error bounds. Again, we will restrict to the upper bound system in which arrivals are rejected when the system is down. Let

$\boldsymbol{F}$: $\lambda(1-\boldsymbol{B})$ the throughput.

$\boldsymbol{Q}$: the mean queue length (or total number of jobs present).

and similarly with subscript U for the upper bound system. The first result, for the throughput (or loss probability), is intuitively obvious and can also be derived analytically by using the comparison (9.4.10) and the product form expression (9.1.2) for the upper bound system. However, it is included to illustrate how result 9.3.5 works out.

The second one is of more practical interest as it is not obvious a priori how much the breakdowns will effect the queue length and how the queue length can be estimated. (Note here that both the lower and upper bound model can intuitively be expected to provide a lower bound for the queue length of the original system).

Result 9.4.7 (Throughput) *With $\boldsymbol{F}_{St}=\boldsymbol{F}_L=\lambda(1-\boldsymbol{B}_{St})$, with $\boldsymbol{B}_{St}$ the standard loss probability of an $M|M|1|N$-queue with traffic load $\rho=[\lambda/\mu]$, and τ the fraction of time that the system is down:*

$$\begin{aligned}&|\boldsymbol{F}-\boldsymbol{F}_U|\leq\tau\boldsymbol{F}_{St}=\\&\tau\rho^N[1-\rho][1-\rho^{N+1}]^{-1}\end{aligned}$$

Proof. To apply result 9.3.5, as in the proof of result 9.4.6 let $\bar{\boldsymbol{q}}$ correspond to the upper bound model. By $\bar{r}(n,\theta)\equiv r(n,\theta)=\mu 1_{(n>0)}1_{(\theta=1)}$, equation (9.4.11) and lemma 9.4.1, condition (9.3.13) is satisfied with

$$\gamma(n,\theta)=\lambda 1_{(n<N)}1_{(\theta=0)}$$

By result 9.3.5 and (9.3.14), we thus find

$$|\boldsymbol{F}-\boldsymbol{F}_U|\leq\sum_{(n,\theta)}\bar{\pi}(n,\theta)\lambda 1_{(n<N)}1_{(\theta=0)}$$

Filling in the product form expression (9.1.2) for $\bar{\pi}(n,\theta)$ as according to the upper bound model, and using its factorizing form in $\pi(n)$ and $\pi(\theta)$ completes the proof.

□

Result 9.4.8 (Queue length) *With $\boldsymbol{Q}_{St}$ the mean queue length of a standard $M|M|1|N$-queue with traffic load $\rho=[\lambda/\mu]$:*

$$|\boldsymbol{Q}-\bar{\boldsymbol{Q}}_U|\leq\tau\boldsymbol{Q}_{St}\left[\frac{\mu}{\mu-\lambda}\right]=\tau\left[\frac{\mu}{\mu-\lambda}\right]\left[\sum_{k=0}^{N}k\rho^k\right]\left[\sum_{k=0}^{N}\rho^k\right]^{-1}$$

Proof. Again, let the $\bar{\boldsymbol{q}}$-system correspond to the upper bound model and $\boldsymbol{q}$ to the original. We need to apply result 9.3.5 and lemma 9.4.2. To this end, first observe that the state spaces of the original and upper bound system are identical. As a consequence, with reference to remark 9.3.9, in order to apply result 9.3.5, we can use condition (9.3.23) instead of condition (9.3.13).

As a consequence, we can reuse (9.4.11) with the bias-terms $\left[V^{k+1}(n+1,\theta)-V^{k}(n,\theta)\right]$ substituted by those from lemma 9.4.2, that is for the upper bound system. Then, by nothing again that $\bar{r}(n,\theta)\equiv r(n,\theta)=n$ to evaluate the mean queue length, by equation (9.4.11) and by lemma 9.4.2, we can verify condition (9.3.23) with

$$\gamma(n,\theta)=\lambda 1_{(n<N)}1_{(\theta=0)}\frac{[1+n]}{[\mu-\lambda]}$$

By result 9.3.5 (with remark 9.3.9 taken in his account), (9.3.14) and by filling in the product form expression (9.1.2) for $\bar{\pi}(n,\theta)$ in (9.3.14), as according to the upper bound model, we find:

$$\begin{aligned}
|\boldsymbol{Q}-\boldsymbol{Q}_U| &\leq \sum\nolimits_{(n,\theta)}\lambda 1_{(n<N)}1_{(\theta=0)}[1+n]\pi(n,\theta)\left[\frac{1}{\mu-\lambda}\right]\\
&=\lambda\tau\left[\frac{1}{\mu-\lambda}\right]\sum\nolimits_{n=0}^{N-1}(1+n)\rho^{n}\left[\sum\nolimits_{k=0}^{N}\rho^{k}\right]^{-1}\\
&=\lambda\tau\left[\frac{1}{\mu-\lambda}\right]\frac{1}{\rho}\left[\sum\nolimits_{k=0}^{N}k\rho^{k}\right]\left[\sum\nolimits_{k=0}^{N}\rho^{k}\right]^{-1}\\
&=\tau\boldsymbol{Q}_{St}\left[\frac{\mu}{\mu-\lambda}\right]
\end{aligned}$$

□

9.5 Application 2: Finite Tandem Queue

9.5.1 Problem Motivation

Reconsider the unsolvable finite tandem queue with blocking as described in section 9.1.2.1, with the single server assumption generalized to service rates $\mu_i(n_i)$ at station i when n_i jobs are present at that station, $i=1,2$. Here the natural assumption is made that $\mu_i(n_i)$ is nondecreasing, for both $i=1,2$. Let $\mu_i(0)=0$, $i=1,2$. Recall the modifications 1 and 2 as described in section 9.1.1.2. We refer to the corresponding systems as:

”lower bound model” (under modification 1) and

”upper bound model” (under modification 2)

as we expect and intend to show:

$$\boldsymbol{B}_L\leq\boldsymbol{B}\leq\boldsymbol{B}_U \tag{9.5.1}$$

with

$\boldsymbol{B}$ the loss probability

$\boldsymbol{B}_L$ the loss probability under modification 1

$\boldsymbol{B}_U$ the loss probability under modification 2

As illustrated already by table 9.2 for the single server case, this ordering result would be of practical interest as the lower and upper bound model exhibit an appealing product form. Indeed, also in this more general case, similar to the detailed (station) balance equations (1.4.1)-(1.4.3) as for the single server case with $\mu_i(n_i)$ substituted when n_i jobs are present, for the upper bound model one directly verifies the product form:

$$\pi_U(n_1,n_2) = c_U \lambda^{n_1+n_2} \left[\prod_{k=1}^{n_1} \mu_1(k)\right]^{-1} \left[\prod_{k=1}^{n_2} \mu_2(k)\right]^{-1} \tag{9.5.2}$$

with c_U the normalizing constant at the set of admissible states:

$$\boldsymbol{S}_U = \{(n_1,n_2) \mid 0 \leq n_1 \leq N_1 \,;\, 0 \leq n_2 \leq N_2 \,;\, n_1+n_2 \neq N_1+N_2\}$$

And similarly, the right hand side of expression (9.5.2) also applies to the steady state distribution $\pi_L(n_1,n_2)$ of the lower bound model, except that c_U has to be replaced by a normalizing constant c_L at

$$\boldsymbol{S}_L = \{(n_1,n_2) \mid n_1 \geq 0 ; n_2 \leq 0 ; n_1+n_2 \leq N_1+N_2\}$$

The loss probabilities $\boldsymbol{B}_L$ and $\boldsymbol{B}_U$ are thus easily computed by:

$$\boldsymbol{B}_L = \pi_L(n_1+n_2 = N_1+N_2) \text{ and}$$
$$\boldsymbol{B}_U = \pi_U(n_1 = N_1 \text{ or } n_2 = N_2)$$

Table 9.4 below gives some more numerical support for the pure multi-server case with N_i servers at station i, $i = 1,2$, and $\rho_1 = \rho_2 = \lambda/\mu_1 = \lambda/\mu_2$ with μ_i the exponential service parameters at station i. A formal proof for (9.5.1) is thus of interest.

Unfortunately, as shown and concluded in section 9.2.3, the technique of "stochastic comparison" by ordering properties of the one-step transition operators necessarily fails for proving the ordering as required for (9.5.1)

In section 9.5.2, therefore, we aim to prove (9.5.1) by the Markov reward approach, that is by result 9.3.2, as based upon technical lemma's for bounding the bias-terms, which are presented in section 9.5.3.

Remark 9.5.1 (Insensitive bounds) *For the pure multi-server case with N_1 and N_2 servers, or for specific disciplines such as a Last-in-First-out preemptive or processor sharing discipline for the single server case, the product form expression (9.5.2) is also insensitive, i.e. the exponentiality assumptions are not required. As a consequence, in these cases one may also expect that the values $\boldsymbol{B}_L$ and $\boldsymbol{B}_U$ provide 'insensitive' bounds, i.e. also apply as bounds for arbitrary service distributions.*

Table 9.4: Loss bounds for pure multi-server-case

ρ_1	ρ_2	N_1	N_2	$\boldsymbol{B}_L$	$\boldsymbol{B}_U$
1	1	4	4	0.001	0.030
10	10	10	10	0.199	0.353
10	10	13	13	0.051	0.156
10	10	15	15	0.012	0.070
10	10	20	20	0.000	0.004

For the pure multi-server case this is indeed proven in [11] in a more technical manner than in this section.

9.5.2 Comparison Result (Bounds)

By virtue of the throughput ($\boldsymbol{F}$) relation $\boldsymbol{F} = \lambda(1-\boldsymbol{B})$, and similarly for the lower and upper bound model, it suffices to show that

$$\boldsymbol{F}_U \leq \boldsymbol{F} \leq \boldsymbol{F}_L \tag{9.5.3}$$

Here the values $\boldsymbol{F}$, $\boldsymbol{F}_L$ and $\boldsymbol{F}_U$ represent the throughputs of the original, lower bound and upper nound model respectively. Furthermore, note that

$$\boldsymbol{S}_U \subseteq \boldsymbol{S} \subseteq \boldsymbol{S}_L \tag{9.5.4}$$

with $\boldsymbol{S}_U \neq \boldsymbol{S} \neq \boldsymbol{S}_L$ and $\boldsymbol{S}$ the state space of the original model:

$$\boldsymbol{S} = \{(n_1,n_2) \mid 0 \leq n_1 \leq N_1; 0 \leq n_2 \leq N_2\} \tag{9.5.5}$$

Let $\boldsymbol{q}$, $\boldsymbol{q}_L$ and $\boldsymbol{q}_U$ be the transition rates and $\boldsymbol{V}^k$, $\boldsymbol{V}^k_L$ and $\boldsymbol{V}^k_U$ the corresponding cumulative reward functions with reward rates r, r_L and r_U as according to the notation in sections 9.2.1 and 9.3.1 for the original, lower and upper bound tandem model respectively. To prove (9.5.3) we can choose:

$$\begin{cases} r(n_1,n_2) = r_L(n_1,n_2) = \mu_2(n_2) \\ r_U(n_1,n_2) = \mu_2(n_2) 1_{n_1<N_1} \end{cases} \tag{9.5.6}$$

In order to apply result 9.3.2 for the upper bound model, with (...) and (9.5.5) taken into account, we need to investigate condition (9.3.7) with $\bar{r} = r_U$ and $\bar{\boldsymbol{q}} = \boldsymbol{q}_U$, for any $(n_1,n_2) \in \boldsymbol{S}_U$. Condition (9.3.7) then leads to the expression:

$$\begin{aligned}
&[r_U(n_1,n_2)-r(n_1,n_2)]+\\
&\textstyle\sum_{(n_1,n_2)'}\left[\boldsymbol{q}_U\left((n_1,n_2),(n_1,n_2)'\right)-\boldsymbol{q}\left((n_1,n_2),(n_1,n_2)'\right)\right]\cdot\\
&\qquad\qquad\qquad\left[\boldsymbol{V}^k\left((n_1,n_2)'\right)-\boldsymbol{V}^k\left((n_1,n_2)\right)\right]\\
&=\\
&-\mu_2(n_2)1_{(n_2=N_2)}\\
&+\lambda 1_{(n_2=N_2)}1_{(n_1+1\leq N_1)}\left[\boldsymbol{V}^k\left((n_1,n_2)\right)-\boldsymbol{V}^k\left((n_1+1,n_2)\right)\right]\\
&+\mu_2(n_2)1_{(n_1=N_1)}1_{(n_2>0)}\left[\boldsymbol{V}^k\left((n_1,n_2)\right)-\boldsymbol{V}^k\left((n_1,n_2-1)\right)\right]
\end{aligned} \tag{9.5.7}$$

By lemma 9.5.2 below, the bias-terms in the second and third term in the right hand side of (9.5.7) are nón-positive, so that this right hand side can be estimated from above by: ≤ 0. By result 9.3.2 (with $\leq$ sign), $\bar{\boldsymbol{G}}=\boldsymbol{F}_U$ and $\boldsymbol{G}=\boldsymbol{F}$, this proves $\boldsymbol{F}_U\leq\boldsymbol{F}$.

In order to apply result 9.3.2 for the lower bound model, as $\boldsymbol{S}\subseteq\boldsymbol{S}_L$, we need to let $\bar{\boldsymbol{q}}$ have the role of the (smaller) original model and $\boldsymbol{q}$ of the lower bound model. For any $(n_1,n_2)\in\bar{\boldsymbol{S}}$ and with $r=r_L$, condition (9.3.7) then leads to:

$$\begin{aligned}
&[r(n_1,n_2)-r_L(n_1,n_2)]+\\
&\textstyle\sum_{(n_1,n_2)'}\left[\boldsymbol{q}\left((n_1,n_2),(n_1,n_2)'\right)-\boldsymbol{q}_L\left((n_1,n_2),(n_1,n_2)'\right)\right]\cdot\\
&\qquad\qquad\qquad\left[\boldsymbol{V}_L^k\left((n_1,n_2)'\right)-\boldsymbol{V}_L^k\left((n_1,n_2)\right)\right]\\
&=\\
&\lambda 1_{(n_1=N_1)}\left[\boldsymbol{V}_L^k\left((n_1,n_2)\right)-\boldsymbol{V}_L^k\left((n_1+1,n_2)\right)\right]+\\
&\mu_1(n_1)1_{(n_2=N_2)}1_{(n_1>0)}\left[\boldsymbol{V}_L^k\left((n_1,n_2)\right)-\boldsymbol{V}_L^k\left((n_1-1,n_2+1)\right)\right]
\end{aligned} \tag{9.5.8}$$

By lemma 9.5.4 below, the bias-terms in the right hand side of (9.5.8) are nón-positive, so that the right hand side can be estimated from above by: ≤ 0. By result 9.3.2 (with $\leq$ sign), $\bar{\boldsymbol{G}}=\boldsymbol{F}$ and $\boldsymbol{G}=\boldsymbol{F}_L$ this proves $\boldsymbol{F}\leq\boldsymbol{F}_L$. The proof of (9.5.1) is hereby completed.

9.5.3 Technical verification of Bias-Terms

Lemma 9.5.2 (Bias-terms for original model) *For all $k\geq 0$, all $(n_1,n_2)\in\boldsymbol{S}$ and with $(n_1+1,n_2)\in\boldsymbol{S}$ in (9.5.9), $(n_1,n_2+1)\in\boldsymbol{S}$ in (9.5.10) and $(n_1-1,n_2+1)\in\boldsymbol{S}$ in (9.5.11):*

$$0\leq\Delta_1\boldsymbol{V}^k(n_1,n_2)=\boldsymbol{V}^k(n_1+1,n_2)-\boldsymbol{V}^k(n_1,n_2)\leq 1 \tag{9.5.9}$$

$$0\leq\Delta_2\boldsymbol{V}^k(n_1,n_2)=\boldsymbol{V}^k(n_1,n_2+1)-\boldsymbol{V}^k(n_1,n_2)\leq 1 \tag{9.5.10}$$

$$0\leq\Delta_3\boldsymbol{V}^k(n_1,n_2)=\boldsymbol{V}^k(n_1-1,n_2+1)-\boldsymbol{V}^k(n_1,n_2)\leq 1 \tag{9.5.11}$$

Proof. As in the proof of lemma 9.4.1, the proof will be given by induction in k. (9.5.9)-(9.5.11) hold for $k = 0$ as $V^0(\cdot,\cdot) \equiv 0$. Suppose that (9.5.9)-(9.5.11) hold for $k \leq m$. We will separately prove (9.5.9), (9.5.10) and (9.5.11) for $k = m+1$.

Before doing so, it is stated that the bias-term equations (9.5.12), (9.5.13) and (9.5.14) can be derived by similar steps as for the derivation of (9.4.4) in section 9.4, either analytically by writing out the reward relations (9.3.5) and collecting terms after substraction, or by probabilistic interpretation by comparing each (possibly after also adding a 'dummy' transition) transition, that can take place in either of the two states, in a pairwise manner. In these equations also terms do appear that are equal to 0 but left in for charity of its (probabilistic) derivation and a possible compensation argument later on (see proof of (9.5.10) for $k = m+1$).

Proof of (9.5.9) for *k=m+1*. By comparing the reward relation (9.3.5) in states (n_1+1, n_2) and (n_1, n_2) with $n_1 < N_1$ and noting that $r(n_1+1, n_2) = r(n_1, n_2)$, as in (9.4.4) we derive:

$$\begin{aligned}
&\Delta_1 V^{m+1}(n_1,n_2)\\
&\quad = h\lambda 1_{(n_1+1<N_1)} \Delta_1 V^m(n_1+1,n_2)\\
&\quad + h\lambda 1_{(n_1+1=N_1)} [V^m(N_1,n_2) - V^m(N_1,n_2)]\\
&\quad + h\mu_1(n_1) 1_{(n_1>0)} 1_{(n_2<N_2)} \Delta_1 V^m(n_1-1,n_2+1)\\
&\quad + h\mu_2(n_2) 1_{(n_2>0)} \Delta_1 V^m(n_1,n_2-1)\\
&\quad + h[\mu_1(n_1+1) - \mu_1(n_1)] 1_{(n_2<N_2)} [V^m(n_1,n_2+1) - V^m(n_1,n_2)]\\
&\quad + \left[1 - h\lambda - h\mu_1(n_1+1) 1_{(n_2<N_2)} - h\mu_2(n_2)\right] \Delta_1 V^m(n_1,n_2)
\end{aligned} \tag{9.5.12}$$

where we note that the second term in the right hand side is equal to 0 while the bias-term in the fifth can be transformed into $\Delta_2 V^m(n_1,n_2)$. By substituting the induction hypotheses $\Delta_1 V^m(\cdot,\cdot) \geq 0$ and $\Delta_2 V^m(\cdot,\cdot) \geq 0$ and noting that $\mu_1(\cdot)$ is nondecreasing, by (9.5.12) we directly verify
$\Delta_1 V^{m+1}(n_1,n_2) \geq 0$.

By substituting the induction hypotheses $\Delta_1 V^m(\cdot,\cdot) \leq 1$ and $\Delta_2 V^m(\cdot,\cdot) \leq 1$, leaving out the 0-term and noting that all coefficients, which all represent probabilities, add up to 1, we can estimate the right hand side from above by 1, i.e.: $\Delta_1^{m+1}(n_1,n_2) \leq 1$. We have thus shown that (9.5.9) also holds for $k = m+1$.

Proof of (9.5.10) for *k=m+1*. Similarly, with $n_2 < N_2$ and by noting that
$r(n_1,n_2+1) - r(n_1,n_2) = \mu_2(n_2+1) - \mu_2(n_2)$, we find:

$$\begin{aligned}
\Delta_2 V^{m+1}&(n_1,n_2+1)\\
&= h[\mu_2(n_2+1)-\mu_2(n_2)]\\
&+ h\lambda 1_{(n_1<N_1)}\Delta_2 V^m(n_1+1,n_2)\\
&+ h\mu_1(n_1)1_{(n_1>0)}1_{(n_2+1<N_2)}\Delta_2 V^m(n_1-1,n_2+1)\\
&+ h\mu_1(n_1)1_{(n_1>0)}1_{(n_2+1=N_2)}\left[V^m(n_1,n_2+1)-V^m(n_1-1,n_2+1)\right]\\
&+ h\mu_2(n_2)1_{(n_2>0)}\Delta_2 V^m(n_1,n_2-1)\\
&+ h[\mu_2(n_2+1)-\mu_2(n_2)]\left[V^m(n_1,n_2)-V^m(n_1,n_2)\right]\\
&+ \left[1-h\lambda 1_{(n_1<N_1)}-h\mu_1(n_1)-h\mu_2(n_2+1)\right]\Delta_2 V^m(n_1,n_2)
\end{aligned} \tag{9.5.13}$$

where we note that the sixth term in the right hand side of (9.5.13) is equal to 0 while the bias-term in the fourth can be transformed into $\Delta_{\mathbf{1}} V^m(n_1-1,n_2+1)$. Clearly, by using that $\mu_2(\cdot)$ is non-decreasing and substituting $\Delta_2 V^m(\cdot,\cdot)\geq 0$ and $\Delta_1 V^m(\cdot,\cdot)\geq 0$ as by hypotheses, by (9.5.13) we directly verify $\Delta_2 V^{m+1}(n_1,n_2)\geq 0$.

To estimate the right hand side of (9.5.13) from above, now we can use that its fifth term is equal to 0, which coefficient equals the additional first reward term $h[\mu_2(n_2+1)-\mu_2(n_2)]$. As a result, by using again that all coefficients sum up to 1 and substituting the induction hypotheses $\Delta_2 V^m(\cdot,\cdot)\leq 1$ and $\Delta_1 V^m(\cdot,\cdot)\leq 1$, we can estimate the right hand side from above by 1, i.e.: $\Delta_2 V^{m+1}(n_1,n_2)\leq 1$. This proves (9.5.10) for $k=m+1$.

Proof of (9.5.11) for *k=m+1*. Similarly for $n_1\leq N_1$ and $n_2<N_2$ we find:

$$\begin{aligned}
\Delta_3 V^{m+1}&(n_1,n_2)\\
&= h[\mu_2(n_2+1)-\mu_2(n_2)]\\
&+ h\lambda 1_{(n_1<N_1)}\Delta_3 V^m(n_1+1,n_2)\\
&+ h\lambda 1_{(n_1=N_1)}\left[\Delta_{\mathbf{2}} V^m(N_1,n_2)\right]\\
&+ h\mu_1(n_1-1)1_{(n_1-1>0)}\Delta_3 V^m(n_1-1,n_2)\\
&+ h\mu_2(n_2)1_{(n_2>0)}\Delta_3 V^m(n_1,n_2-1)\\
&+ h[\mu_1(n_1)-\mu_1(n_1-1)]\left[V^m(n_1-1,n_2+1)-V^m(n_1-1,n_2+1)\right]\\
&+ h[\mu_2(n_2+1)-\mu_2(n_2)]\left[V^m(n_1-1,n_2)-V^m(n_1,n_2)\right]\\
&+ [1-h\lambda-h\mu_1(n_1)-h\mu_2(n_2+1)]\Delta_3 V^m(n_1,n_2)
\end{aligned} \tag{9.5.14}$$

where we first note again that the sixth term in the right hand side is equal to 0. Now note that the seventh term is equal to:

$$h[\mu_2(n_2+1)-\mu_2(n_2)][-\Delta_{\mathbf{1}} V^m(n_1-1,n_2)]$$

and thus nón-positive. However, due to the boundedness hypothesis: $\Delta_1 V^m(n_1-1,n_2)\leq 1$, the first and this seventh term together are still estimated from below by 0. By leaving out the 0-term and substituting $\Delta_2 \boldsymbol{V}^m(\cdot,\cdot) \geq 0$ and $\Delta_3 \boldsymbol{V}^m(\cdot,\cdot) \geq 0$, by (9.5.14) we thus conclude: $\Delta_3 \boldsymbol{V}^{m+1}(n_1,n_2) \geq 0$.

Conversely, by substituting $\Delta_2 \boldsymbol{V}^m(\cdot,\cdot) \leq 1$ and $\Delta_3 \boldsymbol{V}^m(\cdot,\cdot) \leq 1$ and leaving out this nón-positive seventh term to compensate for the additional first term, we can estimate the right hand side of (9.5.14) from above by 1. (As could also be concluded, alternatively, by combining the upper bound 1 from (9.5.10) with the lower bound 0 from (9.5.9) as proven already for k=m+1). We have thus proven (9.5.11) for $k = m+1$.

By induction the proof of lemma 9.5.2 is now completed. □

Remark 9.5.3 *We note that (9.5.11) is not required for substitution within (9.5.7) so as to prove the upper bound $\boldsymbol{B}_U$. The same remark also applies to the ≤ 1 inequalities in (9.5.9)-(9.5.11). However, for instructiveness and completeness they are included, as these type of bias-term estimates will become necessary in lemma 9.5.4 below. In addition, the estimates could be used to also conclude an error bound for the accuracy of the bound as by (9.5.7) and result 9.3.5. However, in this section we aim to restrict to the comparison result (9.5.1), as of sufficient interest by itself.*

Lemma 9.5.4 (Bias-term for lower bound model) *Let the difference terms $\Delta_i \boldsymbol{V}_L^k(n_1,n_2)$, for $i = 1,2,3$, be defined as in lemma 9.5.2, except that the functions $\boldsymbol{V}^k$ are replaced by the functions $\boldsymbol{V}_L^k$ at $\boldsymbol{S}_L$ as by (9.3.7) with $\boldsymbol{q}_L$ substituted for $\boldsymbol{q}$. Then, for all $k \geq 0$:*

$$0 \leq \Delta_1 \boldsymbol{V}_L^k(n_1,n_2) \leq 1 \qquad (n_1+n_2+1 \leq N_1+N_2) \tag{9.5.15}$$

$$0 \leq \Delta_2 \boldsymbol{V}_L^k(n_1,n_2) \leq 1 \qquad (n_1+n_2+1 \leq N_1+N_2) \tag{9.5.16}$$

$$0 \leq \Delta_3 \boldsymbol{V}_L^k(n_1,n_2) \leq 1 \qquad (n_1+n_2 \leq N_1+N_2) \tag{9.5.17}$$

Proof. This will follow similarly to that of lemma 9.5.2 by induction in k. Nevertheless, as the technicalities will appear to be slightly but also essentially different (also see remark 9.5.5 below), the relations still have to be written out in detail below. (Herein, only one 0-term is left in (the sixth in the right hand side of (9.5.19) as it is required for a compensation argument to compensate the reward term).

$$\begin{aligned}
&\Delta_1 \boldsymbol{V}_L^{m+1}(n_1,n_2)\\
&\quad = h\lambda 1_{(n_1+n_2+1<N_1+N_2)} \Delta_1 \boldsymbol{V}_L^m(n_1,n_2)\\
&\quad + h\mu_1(n_1) 1_{(n_1>0)} \Delta_1 \boldsymbol{V}_L^m(n_1-1,n_2+1)\\
&\quad + h\mu_2(n_2) 1_{(n_2>0)} \Delta_1 \boldsymbol{V}_L^m(n_1,n_2-1)\\
&\quad + h[\mu_1(n_1+1)-\mu_1(n_1)] \boldsymbol{\Delta_2} \boldsymbol{V}_L^m(n_1,n_2)\\
&\quad + [1-h\lambda-h\mu_1(n_1+1)] \Delta_1 \boldsymbol{V}_L^m(n_1,n_2)
\end{aligned} \tag{9.5.18}$$

$$\begin{aligned}
&\Delta_2 \boldsymbol{V}_L^{m+1}(n_1,n_2)\\
&\quad = h[\mu_2(n_2+1)-\mu_2(n_2)]\\
&\quad + h\lambda 1_{(n_1+n_2+1<N_1+N_2)}\Delta_2 \boldsymbol{V}_L^m(n_1+1,n_2)\\
&\quad + h\lambda 1_{(n_1+n_2+1=N_1+N_2)}\boldsymbol{\Delta_3} \boldsymbol{V}_L^m(n_1+1,n_2)\\
&\quad + h\mu_1(n_1)1_{(n_1>0)}\Delta_2 \boldsymbol{V}_L^m(n_1-1,n_2)\\
&\quad + h\mu_2(n_2)1_{(n_2>0)}\Delta_2 \boldsymbol{V}_L^m(n_1,n_2-1)\\
&\quad + h[\mu_2(n_2+1)-\mu_2(n_2)]\,[\boldsymbol{V}_L^m(n_1,n_2)-\boldsymbol{V}_L^m(n_1,n_2)]\\
&\quad + [1-h\lambda-h\mu_1(n_1)-h\mu_2(n_2+1)]\,\Delta_2 \boldsymbol{V}_L^m(n_1,n_2)
\end{aligned} \tag{9.5.19}$$

and

$$\begin{aligned}
&\Delta_3 \boldsymbol{V}_L^{m+1}(n_1,n_2)\\
&\quad = h[\mu_2(n_2+1)-\mu_2(n_2)]\\
&\quad + h\lambda 1_{(n_1+n_2+1<N_1+N_2)}\Delta_3 \boldsymbol{V}_L^m(n_1+1,n_2)\\
&\quad + h\mu_1(n_1-1)1_{(n_1-1>0)}\Delta_3 \boldsymbol{V}_L^m(n_1-1,n_2)\\
&\quad + h\mu_2(n_2)1_{(n_2>0)}\Delta_3 \boldsymbol{V}_L^m(n_1,n_2-1)\\
&\quad + h[\mu_2(n_2+1)-\mu_2(n_2)]\,[-\boldsymbol{\Delta_1} \boldsymbol{V}_L^m(n_1-1,n_2)]\\
&\quad + [1-h\lambda-h\mu_1(n_1)-h\mu_2(n_2+1)]\,\Delta_3 \boldsymbol{V}_L^m(n_1,n_2)
\end{aligned} \tag{9.5.20}$$

By induction and the detailed arguments similar to those for (9.5.12)-(9.5.14) in the proof of lemma 9.5.2, the inequalities (9.5.15)-(9.5.17) can now be proven. □

Remark 9.5.5 (Necessity of upper estimates and all bias-terms) *In contrast with lemma 9.5.2 for proving the upper bound $\boldsymbol{B}_U$, now note that the inequality estimate $\Delta_3 \boldsymbol{V}_L^m \geq 0$ is required for proving the lower bound $\boldsymbol{B}_L$, as by (9.5.18). By relation (9.5.20) in turn, and the compensation argument for the nón-positive fifth term in the right hand side of (9.5.20), this necessarily requires an upper estimate $\Delta_1 \boldsymbol{V}_L^m(\cdot,\cdot) \leq 1$. By (9.5.18) in turn this also requires (both a lower and) an upper estimate for $\Delta_2 \boldsymbol{V}_L^m(\cdot,\cdot)$ and by (9.5.19) for $\Delta_3 \boldsymbol{V}_L^m(\cdot,\cdot)$.*

9.6 Application 3: Truncation of Finite Jackson Network

In this section we will apply result 9.3.10 to derive error bound expressions for the truncation of a Finite Jackson Network. A first crucial step is to find bounds for the bias-terms $[\boldsymbol{V}^k(j)-\boldsymbol{V}^k(i)]$ for $\boldsymbol{q}(i,j)>0$, uniformly in all k.

As before, this will be established by inductively exploiting the dynamic reward relation (9.3.5) and the appealing transition structure of queueing networks of the form (9.3.19), as shown in section 9.6.2. First, in section 9.6.1 a more precise description of the FJN of interest will be provided.

9.6.1 Description and motivation.

Consider an open Jackson network with J service stations, numbered $1,\ldots,J$ and Poisson arrival rates λ_i at station i. Let $\lambda = \lambda_1 + \ldots + \lambda_J$ and $p_{0i} = \lambda_i/\lambda$. Upon service completion at station i a job routes to station j with probability p_{ij} or leaves the system with probability $p_{i0} = [1-(p_{i1}+\ldots+p_{iJ})]$ The natural assumption is made that the routing matrix, node 0 included, is irreducible.

Station i has an exponential service rate $\mu_i(n_i)$ when n_i jobs are present. The natural assumption is made that $\mu_i(n_i)$ is non-decreasing. (This will be used in the proof of lemma 9.6.3). Station i has a capacity constraint for no more than N_i jobs. When station i is saturated $(n_i = N_i)$, a job requesting service at station i (arriving from outside or from another node) is lost (i.e. it clears the system) (loss protocol). This loss protocol is motivated by the following application of present day interest.

Special motivation: Mobile Communication Networks

In simplest form, an exponential mobile communication network can be described as follows:

> Calls arrive in a cell j at some arrival rate λ_j (fresh calls). A call duration is assumed to be exponential with parameter μ. A call residing in cell j will move to (another) neighbouring cell k at a rate λ_{jk} (a so called handover call). Within a cell a call requires a frequency channel that is not used by another call within that cell (a free channel). Each cell has a finite number of frequency channels, say N_i in cell i, $i = 1,\ldots,J$. Neighbouring cells cannot have the same frequency channels. When a fresh call cannot find a free channel it is lost. When a handover call cannot find a free new channel in the cell that it is moving to, it is broken off and also lost.
>
> The exponential mobile communication network can directly be reformulated as a Finite Jackson Network by identifying cells with stations and setting:
>
> $$\begin{cases} \mu_i(n_i) = n_i\mu_i \text{ with } \mu_i = [\mu + \sum_j \lambda_{ij}] & \text{(holding or service rate in cell } i) \\ p_{ij} = \lambda_{ij}/\mu_i & \text{(handover probability from } i \text{ to } j) \\ p_{i0} = \mu/\mu_i & \text{(call completion probability at call } i) \end{cases}$$

Remark 9.6.1 (Blocking protocol) *For this application the loss protocol is the natural protocol. As another ‘blocking’ protocol, blocked jobs could be recycled to their originating node.*

Under fairly general conditions this recycling (or repeat) protocol as well as the ’production’ protocol can be shown to be equivalent to the ‘loss’ protocol (also known as communication protocol). (e.g. [48]). Similar results as obtained in section 9.6.3 can also be expected for these protocols.

The finite Jackson network under investigation is generally intractable except in other special cases such as with a reversible routing (e.g. [40], [37]), or with special service or special routing protocols ([15], [37]).

As numerical computations may therefore be executed (either on an exact or approximate basis) (e.g. [36], [32], [60]), a truncation of the state space becomes of interest to limit the computational effort.

9.6.2 Truncation

Let us consider the truncation by restricting the queue lengths to $L_i \leq N_i$ for each station i and assume that the truncated model also operates under the loss protocol. We aim to investigate the consequence of this truncation for the total throughput. Intuitively, at least it seems obvious that the throughput will be reduced by truncation. Nevertheless, as in section 9.4.2, at sample path basis, one can provide counterintuitive examples, as shown by example 9.6.2.

Even a comparison result therefore is still of interest. In fact, in this section we will focus on an error bound, which provides a comparison result at the same time.

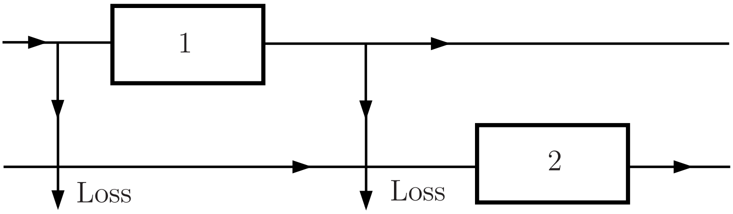


Fig. 9.6: Network Example

Example 9.6.2 (Counterintuitive comparison example) *Consider the example of an original system in figure 9.6 with $N = 2$, $N_1 = 2$, two severs at station 1, $N_2 = 1$ and one server at station 2, $p_{10} = 1/2$ and $p_{12} = 1/2$ and its truncated version with N_1 reduced to $L_1 = 1$. One may intuitively expect that the throughput of the truncated system will be smaller. Consider a sequence of arrivals at station 1 and 2, as shown in figure 9.7, where the second job at station 1 routes to station 2 after its service completion at station 1.*

As shown by figure 9.7, the throughput (accepted number of jobs or successful number of completions) of the truncated system, in this sample path example, appears to be larger.

To apply the results from section 9.3.3, we identify a state i with the queue length vector $\boldsymbol{n} = (n_1, n_2, \ldots, n_J)$ denoting the number of jobs n_i at each station $i = 1, 2, \ldots, J$. All notation from sections 9.2 and 9.3 is adopted accordingly. By $\boldsymbol{e}_i$ we denote the unit vector with the i^{th} component equal to 1, i.e.: $\boldsymbol{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0)$.

Hence by $\boldsymbol{n} - \boldsymbol{e}_i + \boldsymbol{e}_j$ we denote the state with one more job at station j and one less at station i. Similarly we use the notation $\boldsymbol{n} + \boldsymbol{e}_i$ and $\boldsymbol{n} - \boldsymbol{e}_i$. With this notation, the truncation is specified by,

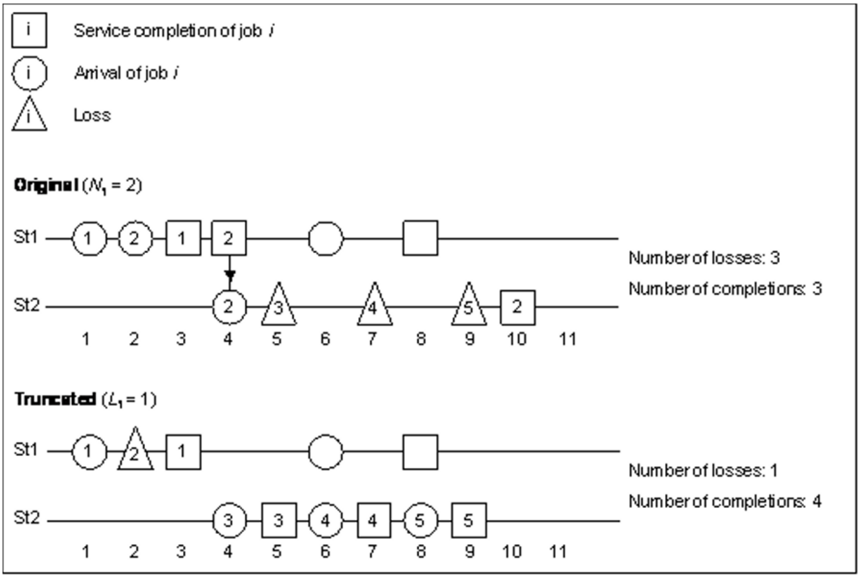


Fig. 9.7: Comparison of original and truncated system

$$\begin{cases} \boldsymbol{S} = \boldsymbol{S}_N = \{\boldsymbol{n} \,|\, n_i \leq N_i & i = 1,\ldots,J\} \\ \bar{\boldsymbol{S}} = \boldsymbol{S}_L = \{\boldsymbol{n} \,|\, n_i \leq L_i \leq N_i & i = 1,\ldots,J\} \end{cases}$$

and for all $(i, j = 1,\ldots,J)$:

$$\begin{cases} \boldsymbol{t}[\boldsymbol{n},\boldsymbol{n}+\boldsymbol{e}_j] & = \boldsymbol{n} & \text{for } n_j = L_j \text{ and } \boldsymbol{n} \in \bar{\boldsymbol{S}} \\ \boldsymbol{t}[\boldsymbol{n},\boldsymbol{n}-\boldsymbol{e}_i+\boldsymbol{e}_j] & = \boldsymbol{n}-\boldsymbol{e}_i & \text{for } n_j = L_j \text{ and } \boldsymbol{n} \in \bar{\boldsymbol{S}} \end{cases} \tag{9.6.1}$$

To apply result 9.3.2, for the left hand side of inequality (9.3.25) we obtain:

$$\begin{aligned} &\textstyle\sum_{n' \notin \boldsymbol{S}_L} \boldsymbol{q}(\boldsymbol{n},\boldsymbol{n}') \left[\boldsymbol{V}^k(\boldsymbol{n}') - \boldsymbol{V}^k\left(\boldsymbol{t}[\boldsymbol{n},\boldsymbol{n}']\right)\right] = \\ &\textstyle\sum_{j=1,\ldots,J} 1_{(n_j=L_j)} \lambda_j \left[\boldsymbol{V}^k(\boldsymbol{n}+\boldsymbol{e}_j) - \boldsymbol{V}^k(\boldsymbol{n})\right] + \\ &\textstyle\sum_{k=1,\ldots,J} \mu_k(n_k) p_{kj} 1_{(n_j=L_j)} \left[\boldsymbol{V}^k(\boldsymbol{n}-\boldsymbol{e}_k+\boldsymbol{e}_j) - \boldsymbol{V}^k(\boldsymbol{n}-\boldsymbol{e}_k)\right] \end{aligned} \tag{9.6.2}$$

In order to estimate this expression from above and below it thus suffices to estimate bias-terms of the form $\left[\boldsymbol{V}^k(\boldsymbol{n}+\boldsymbol{e}_i) - \boldsymbol{V}^k(n)\right]$. This will be established in lemma 9.6.3 below. Herein, we use the shorthand notation: $\Delta_i f(\boldsymbol{n}) = f(\boldsymbol{n}+\boldsymbol{e}_i) - f(\boldsymbol{n})$. As measure of interest we consider the system throughput $\boldsymbol{F}$ by:

$$\boldsymbol{F} = \sum_{\boldsymbol{n} \in \boldsymbol{S}} \pi(\boldsymbol{n}) r(\boldsymbol{n})$$

with

$$r(\boldsymbol{n}) = \sum_{p} \lambda_p 1_{(n_p < N_p)}$$

Lemma 9.6.3 *For all states $\boldsymbol{n} \in \boldsymbol{S}$, any station l such that $\boldsymbol{n} + \boldsymbol{e}_l \in \boldsymbol{S}$ and all $k \geq 0$:*

$$0 \geq \Delta_l \boldsymbol{V}^k(\boldsymbol{n}) = \boldsymbol{V}^k(\boldsymbol{n} + \mathbf{e}_l) - \boldsymbol{V}^k(\boldsymbol{n}) \geq -1 \tag{9.6.3}$$

Proof. The proof will follow by induction in k. Clearly (9.6.3) holds for $k = 0$. Let (9.6.3) hold for $k = t$ and all $\{\boldsymbol{n}, \boldsymbol{n} + \boldsymbol{e}_l\} \in \boldsymbol{S}_N$. We need to verify (9.6.3) for $k = t + 1$. To this end by writing out (9.3.5) in state $\boldsymbol{n}$, we find:

$$\begin{aligned}
&\boldsymbol{V}^{t+1}(\boldsymbol{n}) \\
&\quad = h \textstyle\sum_p \lambda_p 1_{(n_p < N_p)} \\
&\quad + h \textstyle\sum_j \lambda_j 1_{(n_j < N_j)} \boldsymbol{V}^t(\boldsymbol{n} + \boldsymbol{e}_j) \\
&\quad + h \textstyle\sum_j \lambda_j 1_{(n_j = N_j)} \boldsymbol{V}^t(\boldsymbol{n}) \\
&\quad + h \textstyle\sum_i \mu_i(n_i) 1_{(n_i > 0)} p_{i0} \boldsymbol{V}^t(\boldsymbol{n} - \boldsymbol{e}_i) \\
&\quad + h \textstyle\sum_i \mu_i(n_i) 1_{(n_i > 0)} \sum_j p_{ij} 1_{(n_j < N_j)} \boldsymbol{V}^t(\boldsymbol{n} - \boldsymbol{e}_i + \boldsymbol{e}_j) \\
&\quad + h \textstyle\sum_i \mu_i(n_i) 1_{(n_i > 0)} \sum_j p_{ij} 1_{(n_j = N_j)} \boldsymbol{V}^t(\boldsymbol{n} - \boldsymbol{e}_i) \\
&\quad + \left[1 - h \textstyle\sum_j \lambda_j - h \sum_i \mu_i(n_i)\right] \boldsymbol{V}^t(\boldsymbol{n})
\end{aligned} \tag{9.6.4}$$

and similarly in state $\boldsymbol{n} + \boldsymbol{e}_l$:

$$
\begin{aligned}
&\boldsymbol{V}^{t+1}(\boldsymbol{n}+\boldsymbol{e}_i)\\
&\quad = h\sum\nolimits_{p\neq l}\lambda_p 1_{(n_p<N_p)} + h\lambda_l 1_{(n_l+1<N_1)}\\
&\quad + \sum\nolimits_{j\neq l} h\lambda_j \left\{1_{(n_j<N_j)}\boldsymbol{V}^t(\boldsymbol{n}+\boldsymbol{e}_l+\boldsymbol{e}_j) + 1_{(n_j=N_j)}\boldsymbol{V}^t(\boldsymbol{n}+\boldsymbol{e}_l)\right\}\\
&\quad + h\lambda_l 1_{(n_l+1<N_l)}\boldsymbol{V}^t(\boldsymbol{n}+\boldsymbol{e}_l+\boldsymbol{e}_l) + h\lambda_l 1(n_l+1=N_l\boldsymbol{V}^t(\boldsymbol{n}+\boldsymbol{e}_l))\\
&\quad + h\sum\nolimits_{i\neq l}\mu_i(n_i)p_{i0}\boldsymbol{V}'(\boldsymbol{n}+\boldsymbol{e}_l-e_i)\\
&\quad + h\sum\nolimits_{i\neq l}\mu_i(n_i)\sum\nolimits_{j\neq l}p_{ij}1_{(n_j<N_j)}\boldsymbol{V}^t(\boldsymbol{n}+\boldsymbol{e}_l+\boldsymbol{e}_j-e_i)\\
&\quad + h\sum\nolimits_{i\neq l}\mu_i(n_i)\sum\nolimits_{j\neq l}p_{ij}1_{(n_j=N_j)}\boldsymbol{V}^t(\boldsymbol{n}+\boldsymbol{e}_l-e_i)\\
&\quad + h\sum\nolimits_{i\neq l}\mu_i(n_i)p_{il}1_{(n_l+1<N_l)}\boldsymbol{V}^t(\boldsymbol{n}+\boldsymbol{e}_l+\boldsymbol{e}_l-e_i)\\
&\quad + h\sum\nolimits_{i\neq l}\mu_i(n_i)p_{il}1_{(n_l+1=N_l)}\boldsymbol{V}^t(\boldsymbol{n}+\boldsymbol{e}_l-e_i)\\
&\quad + h\mu_l(n_l)p_{l0}\boldsymbol{V}^t(\boldsymbol{n}) + h\mu_l(n_l)p_{ll}\boldsymbol{V}^t(\boldsymbol{n}+\boldsymbol{e}_l)\\
&\quad + h\mu_l(n_l)h\sum\nolimits_{i\neq l}\left\{1_{(n_j<N_j)}\boldsymbol{V}^t(\boldsymbol{n}+\boldsymbol{e}_j) + 1_{(n_j=N_j)}\boldsymbol{V}^t(\boldsymbol{n})\right\}\\
&\quad + h[\mu_l(n_l+1)-\mu_l(n_l)]p_{l0}\boldsymbol{V}^t(\boldsymbol{n})\\
&\quad + h[\mu_l(n_l+1)-\mu_l(n_l)]p_{ll}\boldsymbol{V}^t(\boldsymbol{n}+\boldsymbol{e}_l)\\
&\quad + h[\mu_l(n_l+1)-\mu_l(n_l)]\sum\nolimits_{i\neq l}p_{lj}\left\{1_{(n_j<N_j)}\boldsymbol{V}^t(\boldsymbol{n}+\boldsymbol{e}_j) + 1_{(n_j=N_j)}\boldsymbol{V}^t(\boldsymbol{n})\right\}\\
&\quad + \left\{1-h\sum\nolimits_j\lambda_j - h\sum\nolimits_{i\neq l}\mu_i(n_i) - h\mu_l(n_l+1)\right\}\boldsymbol{V}^t(\boldsymbol{n}+\boldsymbol{e}_l)
\end{aligned}
\tag{9.6.5}
$$

To subtract (9.6.4) from (9.6.5) and to compare transitions pairwise, make the following modifications in (9.6.4):

- Rewrite the summation for all j in a summation for all $j \neq l$ and its separate expression $j = l$.
- Note that since $n_l + 1 \leq N_l$ also $n_l < N_l$ and rewrite:

$$
\begin{aligned}
&\left[h\lambda_l + h\sum\nolimits_{i\neq l}\mu_i(n_i)p_{il}\right]\boldsymbol{V}^t(\boldsymbol{n}+\boldsymbol{e}_l) =\\
&\left[h\lambda_l + h\sum\nolimits_{i\neq l}\mu_i(n_i)p_{il}\right]\left[1_{(n_l+1<N_l)}\boldsymbol{V}^t(\boldsymbol{n}+\boldsymbol{e}_l) + 1_{(n_l+1=N_l)}\boldsymbol{V}^t(\boldsymbol{n}+\boldsymbol{e}_l)\right]
\end{aligned}
$$

- Artificially add and subtract a departure from station l at a rate $[\mu_l(n_l+1)-\mu_l(n_l)]$ that leaves the state unchanged, that is; artificially add and subtract the expression:

$$
\begin{aligned}
&h\left[\mu_l(n_l+1)-\mu_l(n_l)\right]p_{l0}\boldsymbol{V}^t(\boldsymbol{n})+\\
&h\left[\mu_l(n_l+1)-\mu_l(n_l)\right]p_{ll}\boldsymbol{V}^t(\boldsymbol{n})+\\
&\sum\nolimits_{j\neq l}p_{lj}\left[1_{(n_j<N_j)}\boldsymbol{V}^t(\boldsymbol{n}) + 1_{(n_j<N_j)}\boldsymbol{V}^t(\boldsymbol{n})\right]
\end{aligned}
$$

Then, subtracting (9.6.4) from (9.6.5) finally leads to the following difference expression. Again, some terms will in fact be equal to 0 but left in for clarity of derivation.

$$\begin{aligned}
&\boldsymbol{V}^{t+1}(\boldsymbol{n}+\boldsymbol{e}_i)\\
&\quad = h1_{(n_l+1=N_l)}[-\lambda_l]\\
&\quad + h\textstyle\sum_{j\neq l}\lambda_j 1_{(n_j<N_j)}\Delta_l\boldsymbol{V}^t(\boldsymbol{n}+\boldsymbol{e}_j)\\
&\quad + h\textstyle\sum_{j\neq l}\lambda_j 1_{(n_j=N_j)}\Delta_l\boldsymbol{V}^t(\boldsymbol{n})\\
&\quad + h\lambda_j 1_{(n_j+1<N_j)}\Delta_l\boldsymbol{V}^t(\boldsymbol{n}+\boldsymbol{e}_l)\\
&\quad + h\lambda_j 1_{(n_j+1=N_j)}\left[\boldsymbol{V}^t(\boldsymbol{n}+\boldsymbol{e}_l)-\boldsymbol{V}^t(\boldsymbol{n}+\boldsymbol{e}_l)\right]\\
&\quad + h\textstyle\sum_{i\neq l}\mu_i(n_i)p_{i0}\Delta_l\boldsymbol{V}^t(\boldsymbol{n}-\boldsymbol{e}_i)\\
&\quad + h\textstyle\sum_{i\neq l}\mu_i(n_i)\sum_{j\neq l}p_{ij}1_{(n_j<N_j)}\Delta_l\boldsymbol{V}^t(\boldsymbol{n}+\boldsymbol{e}_j-\boldsymbol{e}_i)\\
&\quad + h\textstyle\sum_{i\neq l}\mu_i(n_i)\sum_{j\neq l}p_{ij}1_{(n_j=N_j)}\Delta_l\boldsymbol{V}^t(\boldsymbol{n}-\boldsymbol{e}_i)\\
&\quad + h\textstyle\sum_{i\neq l}\mu_i(n_i)p_{il}1_{(n_l+1<N_l)}\Delta_l\boldsymbol{V}^t(\boldsymbol{n}+\boldsymbol{e}_j-\boldsymbol{e}_i)\\
&\quad + h\textstyle\sum_{i\neq l}\mu_i(n_i)p_{il}1_{(n_l+1=N_l)}\left[\boldsymbol{V}^t(\boldsymbol{n}+\boldsymbol{e}_l-\boldsymbol{e}_i)-\boldsymbol{V}^t(\boldsymbol{n}+\boldsymbol{e}_l-\boldsymbol{e}_i)\right]\\
&\quad + h\mu_l(n_l)p_{l0}\Delta_l\boldsymbol{V}^t(\boldsymbol{n}-\boldsymbol{e}_l)+h\mu_l(n_l)p_{ll}\Delta_l\boldsymbol{V}^t(\boldsymbol{n})\\
&\quad + h[\mu_l(n_l+1)-\mu_l(n_l)]p_{l0}\left[\boldsymbol{V}^t(\boldsymbol{n})-\boldsymbol{V}^t(\boldsymbol{n})\right]\\
&\quad + h[\mu_l(n_l+1)-\mu_l(n_l)]p_{ll}\Delta_l\boldsymbol{V}^t(\boldsymbol{n})\\
&\quad + h\mu_l(n_l)\textstyle\sum_{j\neq l}p_{lj}\left[1_{(n_j<N_j)}\Delta_l\boldsymbol{V}^t(\boldsymbol{n}+\boldsymbol{e}_j-\boldsymbol{e}_l)+1_{(n_j=N_j)}\Delta_l\boldsymbol{V}^t(\boldsymbol{n}-\boldsymbol{e}_l)\right]\\
&\quad + h[\mu_l(n_l+1)-\mu_l(n_l)]\textstyle\sum_{i\neq l}p_{lj}1_{(n_j<N_j)}\Delta_l\boldsymbol{V}^t(\boldsymbol{n})\\
&\quad + h[\mu_l(n_l+1)-\mu_l(n_l)]\textstyle\sum_{i\neq l}p_{lj}1_{(n_j=N_j)}\left[\boldsymbol{V}^t(\boldsymbol{n})-\boldsymbol{V}^t(\boldsymbol{n})\right]\\
&\quad + \left\{1-h\textstyle\sum_j\lambda_j-h\sum_{i\neq l}\mu_i(n_i)-h\mu_l(n_l+1)\right\}\Delta_l\boldsymbol{V}^t(\boldsymbol{n})
\end{aligned}\tag{9.6.6}$$

By substituting the induction hypothesis: $\Delta_l\boldsymbol{V}^t(\boldsymbol{n})\leq 0$ and $\Delta_j\boldsymbol{V}^t(\boldsymbol{n})\leq 0$ for all j, and deleting the 0 terms in the right hand side of (9.6.6), from (9.6.6) we now directly conclude:

$$\Delta_l\boldsymbol{V}^{t+1}(\boldsymbol{n})\leq 0$$

To estimate the right hand side of (9.6.6) from below, now substitute the induction hypothesis $\Delta_l\boldsymbol{V}^t(\boldsymbol{n})\geq -1$ and $\Delta_j\boldsymbol{V}^t(\boldsymbol{n})\geq -1$ for all j. Furthermore, note that the term with coefficient $h\lambda_l 1_{(n_l+1=N_l)}$ is equal to 0, which compensates for the first extra negative term

$$h1_{(n_l+1=N_l)}[-\lambda_l]$$

By also noting that all coefficients (which in fact represent transition probabilities) sum up to 1, now conclude

$$\Delta_l\boldsymbol{V}^{t+1}(\boldsymbol{n})\geq -1$$

The induction completes the proof. □

By lemma 9.6.3 we can now apply the general truncation result 9.3.10. By combining result 9.3.10, lemma 9.6.3 and expression (9.6.2), this leads to the following result.

Result 9.6.4 (Computational Error Bound) *Consider a FJN with capacity constraints N_i at station i, $i = 1,\dots,J$ and its truncation with capacity constraints $L_i \leq N_i$ at station i, $i = 1,\dots,J$. Let $\boldsymbol{F}_N$ and $\boldsymbol{F}_L$ be the corresponding system throughputs and $\{\pi_L(\boldsymbol{n}) \mid \boldsymbol{n} \in \boldsymbol{S}_L\}$ the steady state distribution of the truncated FJN. Then:*

$$0 \leq \boldsymbol{F}_N - \boldsymbol{F}_L \leq \sum\nolimits_{\boldsymbol{n}} \pi_L(\boldsymbol{n}) \sum\nolimits_{j=1}^{J} 1_{(n_j=L_j)} \left[\lambda_j + \sum\nolimits_k \mu_k(n_k) p_{kj}\right] \tag{9.6.7}$$

Remark 9.6.5 (Comparison and monotonicity results: literature) *Comparison or monotonicity results have been reported explicitly for Jackson Networks but but only under special conditions, as a product form, single servers or infinite capacities, as in [1], [2], [53], [65]. The results in these references rely upon sample path comparison.*

However, as has been shown in example 9.6.2, for finite queueing systems a sample path comparison can be violated (also see [20], [27], [58], [62], [59]). Nevertheless, as shown by result 9.6.4, a comparison result at expectation basis can still be established. This result seems to be new.

Remark 9.6.6 (Computational error bound) *Result 9.6.4 enables one to provide a secure bounding interval:*

$$\boldsymbol{F}_L \leq \boldsymbol{F}_N \leq \boldsymbol{F}_L + \delta_L \tag{9.6.8}$$

where δ_L is the upper estimate from (9.6.7), once we have computed the distribution $\{\pi_L(\boldsymbol{n})\}$ for the truncated system. It could therefore be referred to as a computational error bound.

9.6.3 Analytic Error bound

As the error bound in (9.6.7), that is δ_L in (9.6.8), might still be computationally complicated it is more appealing to replace δ_L by an analytic expression. As will be shown below, this can be established by comparing the truncated system with an infinite system that exhibits a product form. This will lead to an analytic bound $\delta_L \leq \delta_\infty$.

Consider the infinite Jackson network which allows an infinite queue length at station i with service rate:

$$s_i(n_i) = \begin{cases} \mu_i(n_i) & (n_i \leq L_i) \\ \max\left[\mu_i(L_i), n_i\mu_i\right] & (n_i > L_i) \end{cases}$$

$i = 1,\dots,J$. As justified by a natural irreducibility assumption, let
$\{v_i\,; i = 1,\dots,J\}$ be the unique solution of the traffic equations:

$$v_i = \lambda_i + \sum\nolimits_k v_k p_{ki} \qquad (i = 1,\dots,J)$$

The corresponding steady state distribution, denoted by $\pi_\infty(n)$ at $\boldsymbol{S}_\infty = \{\boldsymbol{n} \mid n_i \geq 0, i = 1,\dots,J\}$ then exhibits the product form:

$$\pi_\infty(\boldsymbol{n}) = \prod_{i=1}^{J} \left\{ c_i \nu_i^{n_i} \left[\prod_{k=1}^{n_i} s_i(k)\right]^{-1} \right\} \qquad (\boldsymbol{n} \in \boldsymbol{S}_\infty) \tag{9.6.9}$$

with c_i a normalizing constant for each station i. Let $\beta(\boldsymbol{n})$ represent the infinite expansion of the boundary (state) as occurring in (9.6.7). More precisely, that is:

$$\beta(\boldsymbol{n}) = \sum_{j=1}^{J} 1_{(n_j \geq L_j)} \left[\lambda_j + \sum_k \mu_k(n_k) p_{kj}\right] \tag{9.6.10}$$

Result 9.6.7 (Analytic Error Bound) *With $\boldsymbol{F}_N$ and $\boldsymbol{F}_L$ as in result 9.6.4 and $\pi_\infty(\boldsymbol{n})$ by (9.6.9):*

$$\boldsymbol{F}_L \leq \boldsymbol{F}_N \leq \boldsymbol{F}_L + \delta_\infty$$

with

$$\delta_\infty = \sum_n \pi_\infty(\boldsymbol{n}) \beta(\boldsymbol{n}) \tag{9.6.11}$$

Proof. The proof follows as immediate consequence of result 9.6.8 once we have shown that the steady-state probabilities $\pi_L(\boldsymbol{n})$ in (9.6.7) at boundary states can be bounded from above by $\pi_\infty(\boldsymbol{n})$ ’tail’ probabilities beyond these boundaries. To this end we can use the comparison results from section 9.2.2, more precisely result 9.2.1. To apply result 9.2.1, in the setting of section 9.2.1, let the infinite Jackson network represent the original and the truncated Jackson network the modified system.

Let $\boldsymbol{P}_L$ and $\boldsymbol{P}_\infty$ be the uniformized transition matrices for the truncation and infinite Jackson network as according to their definition (9.2.3). Let $\boldsymbol{T}_L$ and $\boldsymbol{T}_\infty$ be the corresponding expectation operators.

Let $\mathbb{M}$ represent the class of component-wise monotone functions as defined by

$$\mathbb{M} = \{f : \boldsymbol{S}_\infty \to \mathbb{R} \mid f(\boldsymbol{n} + \boldsymbol{e}_l) - f(\boldsymbol{n}) \geq 0\,;\, \boldsymbol{n} \in \boldsymbol{S}_\infty\,;\, l = 1,\dots,J\} \tag{9.6.12}$$

Now let us first show that the infinite system preserves this monotonicity. That is, condition (9.2.8) as:

$$\boldsymbol{T}_\infty f \in \mathbb{M} \ \text{ for any } f \in \mathbb{M} \tag{9.6.13}$$

Let $f \in \mathbb{M}$. Then similarly to the derivation of (9.6.6) by writing out all transition probabilities in state $(\boldsymbol{n} + \boldsymbol{e}_l)$ and state $\boldsymbol{n}$ where we can make the simplifications that $N_j = \infty$ for all j, and by comparing the transitions in these two states pairwise, along the lines of (9.6.4) and (9.6.5) and after the appropriate substitutions as used for (9.6.6), we obtain:

$$\begin{aligned}
&(\boldsymbol{T}_\infty f)(\boldsymbol{n}+\boldsymbol{e}_l) - \boldsymbol{T}_\infty f(\boldsymbol{n})\\
&\quad = h\sum\nolimits_j \lambda_j\,[f(\boldsymbol{n}+\boldsymbol{e}_j - f(\boldsymbol{n})]\\
&\quad + h\sum\nolimits_i \mu_i(n_i)\sum\nolimits_j p_{ij}\,[f(\boldsymbol{n}+\boldsymbol{e}_l-\boldsymbol{e}_i+\boldsymbol{e}_j) - f(\boldsymbol{n}-\boldsymbol{e}_i+\boldsymbol{e}_j)]\\
&\quad + h\,[\mu_l(\boldsymbol{n}_l+1)-\mu_l(\boldsymbol{n}_l)]\sum\nolimits_j p_{lj}\,[f(\boldsymbol{n},+\boldsymbol{e}_j)-f(\boldsymbol{n})]\\
&\quad + h\sum\nolimits_i \mu_i(\boldsymbol{n}_i)p_{i0}\,[f(\boldsymbol{n}+\boldsymbol{e}_l-\boldsymbol{e}_i)-f(\boldsymbol{n})]\\
&\quad + h\,[\mu_l(\boldsymbol{n}_l+1)-\mu_l(\boldsymbol{n}_l)]\,p_{l0}\,[f(\boldsymbol{n})-f(\boldsymbol{n})]\\
&\quad + \left[h-\sum\nolimits_j\lambda_j - h\sum\nolimits_{i\neq l}\mu_i(n_i)-h\mu_l(n_l+1)\right][f(\boldsymbol{n}+\boldsymbol{e}_l)-f(\boldsymbol{n})]
\end{aligned}\tag{9.6.14}$$

(Here the one but last term is indeed equal to 0 but kept for its clarity of derivation). By using that $f\in\mathbb{M}$, so that we can substitute: $f(\boldsymbol{n}+\boldsymbol{e}_j)-f(\boldsymbol{n})\geq 0$ for all $\boldsymbol{n}$ and j, the right hand side of (9.6.14) is directly estimated from below by 0, for arbitrary $\boldsymbol{n}$ and l. This proves (9.6.13), that is, (9.2.8) from section 9.2.2.

Next, as second step, we need to verify condition (9.2.9) for any $f\in\mathbb{M}$. This however follows directly as for any $f\in\mathbb{M}$ and $\boldsymbol{n}\in\boldsymbol{S}_L$:

$$\begin{aligned}
&(\boldsymbol{T}_L-\boldsymbol{T}_\infty)f(\boldsymbol{n}) =\\
&\sum\nolimits_{j=1}^J 1_{(n_j=L_j)}\lambda_j\,[f(\boldsymbol{n})-f(\boldsymbol{n}+\boldsymbol{e}_j)]+\\
&\sum\nolimits_{k=1}^J \mu_k(\boldsymbol{n}_k)\sum\nolimits_{j=1}^J p_{kj}1_{(n_j=L_j)}\,[f(\boldsymbol{n}-\boldsymbol{e}_k)-f(\boldsymbol{n}+\boldsymbol{e}_j-\boldsymbol{e}_k)]\leq 0
\end{aligned}\tag{9.6.15}$$

As third step, to verify condition (9.2.11), now note that for β as defined by (9.6.10) and $\mathbb{M}$ as defined by (9.6.12):

$$\beta\in\mathbb{M}\tag{9.6.16}$$

By combining (9.6.13), (9.6.15) and (9.6.16) and applying result 9.2.1, we may now conclude:

$$\begin{aligned}
\boldsymbol{G}_L &= \sum\nolimits_{\boldsymbol{n}}\pi_L(\boldsymbol{n})\beta(\boldsymbol{n}) = \delta_L\leq\\
\boldsymbol{G}_\infty &= \sum\nolimits_{\boldsymbol{n}}\pi_\infty(\boldsymbol{n})\beta(\boldsymbol{n}) = \delta_\infty
\end{aligned}\tag{9.6.17}$$

The proof of result 9.6.7 is now completed by applying result 9.6.4 and estimating the right hand side of (9.6.17) from above by:

$$\delta_L\leq\delta_\infty$$

□

9.6.4 Application: Cellular Mobile Network Application

For the special application of a cellular mobile network as in figure 9.8, we can set $\mu_i(n_i) = n_i\mu_i$.

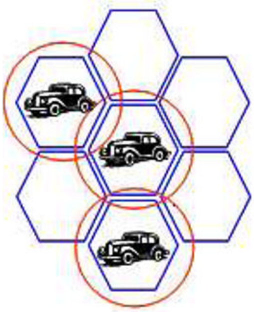


Fig. 9.8: Cellular mobile network ($J = 7$)

In this case the infinite extension and its product form become:

$$\pi_\infty(\boldsymbol{n}) = \prod_{k=1}^{J} \left\{ e^{-\rho_k} \frac{1}{n_k!} [\rho_k]^{n_k} \right\} \quad (\boldsymbol{n} \in \boldsymbol{S}_\infty) \quad \text{with} \quad \pi_k(t) = e^{-\rho_k} \frac{1}{t!} [\rho_k]^t \quad (t \geq 0) \tag{9.6.18}$$

Based upon the decomposability of this expression in individual stations as if they are independent and the traffic equations for $\{\nu_i\}$, we derive:

$$\begin{aligned} \delta_\infty &= \textstyle\sum_n \pi_\infty(\boldsymbol{n}) g(\boldsymbol{n}) = \sum_n \pi_\infty(\boldsymbol{n}) \sum_{j=1}^J 1_{(n_j > L_j)} \left\{ \lambda_j + \sum_{k=1}^J n_k \mu_k p_{kj} \right\} \\ &= \textstyle\sum_{j=1}^J \left[\displaystyle\sum_{n_j = L_j}^{\infty} \pi_j(n_j) \right] \left\{ \lambda_j + \displaystyle\sum_{k \neq j}^{J} \left[\sum_{n_k=0}^{\infty} \pi_k(n_k) n_k \mu_k \right] p_{kj} \right\} \\ &= \textstyle\sum_{j=1}^J \left\{ \lambda_j + \sum_{k\neq j}^J \sum_{t=0}^\infty e^{-\nu_k/\mu_k} \dfrac{t}{t!} \left[\dfrac{\nu_k}{\mu_k} \right]^t \mu_k p_{kj} \right\} \sum_{n_j = L_j}^\infty \pi_j(n_j) \\ &= \textstyle\sum_{j=1}^J \left\{ \lambda_j + \sum_{k=1}^J \nu_k p_{kj} \right\} \sum_{n_j=L_j}^\infty \pi_j(n_j) \\ &= \textstyle\sum_{j=1}^J \nu_j e^{-\rho_j} \sum_{t=L_j}^\infty \dfrac{1}{t!} [\rho_j]^t \end{aligned} \tag{9.6.19}$$

Furthermore, in order to provide a relative error bound $(\boldsymbol{F}_N - \boldsymbol{F}_L)/\boldsymbol{F}_N$ rather than just an absolute error $(\boldsymbol{F}_N - \boldsymbol{F}_L)$ as based upon (9.6.11), by (9.6.17) with L replaced by N and with $f(\boldsymbol{n}) = 1_{(n_j \geq N_j)}$, we can conclude that:

$$\pi_N(n_j = N_j) \leq \pi_\infty(n_j \geq N_j)$$

so that

$$
\begin{aligned}
\boldsymbol{F}_N = \textstyle\sum_{j=1}^{S} \lambda_j \pi_N(n_j < N_j) \geq \textstyle\sum_{j=1}^{J} \lambda_j \pi_\infty(n_j < N_j) = \\
\textstyle\sum_{j=1}^{J} \lambda_j \left\{ 1 - e^{-\rho_j} \textstyle\sum_{t=N_j}^{\infty} \frac{1}{t!} [\rho_j]^t \right\}
\end{aligned}
\tag{9.6.20}
$$

As a consequence, by result 9.6.4, (9.6.19) and (9.6.20), the following relative error bound can now be concluded directly for a channel reduction to L_j channels for cell j, $j = 1, \ldots, J$.

Result 9.6.8 *Let* $\rho_j = [\nu_j / \mu_j]$ *and*

$$
\boldsymbol{B}_j(s) = e^{-\rho_j} \textstyle\sum_{k=s}^{\infty} \frac{1}{k!} [\rho_j]^k \tag{9.6.21}
$$

Then

$$
\Delta = \left[\frac{\boldsymbol{F}_N - \boldsymbol{F}_L}{\boldsymbol{F}_N} \right] \leq \frac{\sum_j \boldsymbol{B}_j(L_j)}{\sum_j \lambda_j [1 - \boldsymbol{B}_j(L_j)]} \tag{9.6.22}
$$

Remark 9.6.9 *Usually, the number of channels in cell j is determined such that a service level of* $\boldsymbol{S}_j \cdot 100\%$ *is guaranteed where* $\boldsymbol{S}_j = 1 - \boldsymbol{B}_j$ *with* $\boldsymbol{B}_j$ *Erlang's loss probability of a multi-server* $M|M|N_j|N_j$ *loss system with* N_j *servers and traffic intensity* $\rho_j = [\nu_j / \mu_j]$*, as if in isolation. Such first order approximations have been used to establish fixed point approximations (e.g. [41], [49]).*

Remark 9.6.10 (Markov Reward Approach and stochastic comparison combination) *Note that the analytic error bound results 9.6.7 and 9.6.8 are essentially based on the error bound result 9.6.4, and thus the general error bound result 9.3.10 and 9.3.5, as well as the stochastic comparison result 9.2.1. Also the combination of the two approaches, the Markov reward approach and stochastic comparison, thus appears to be fruitful.*

9.7 Evaluation

In this chapter the Markov reward approach has been discussed in order to compare two related queueing networks, where one may typically be thought of as a modification of the other for computational simplification. This approach has both advantages and disadvantages as opposed to the more standard stochastic comparison approach, most notably among which, as advantages:

- It may also lead to (analytic) error bounds for the discrepancy
- It may still apply while stochastic comparison fails,

while as disadvantages:

- It requires exponentiality assumptions
- It can technically be more complicated.

Also a combination of both approaches might become useful to secure simpler analytic bounds for the error bounds. A number of extensions as well as questions are still open for further research.

9.7.1 Extensions

1. **Non-exponentiality.** In principle non-exponential queueing networks can be covered by the MRA by using phase-type distributions, possibly in combination with weak convergence arguments, to approximate 'arbitrary' non-exponential service and interarrival distributions. However, the notational extensions and technical verifications of the necessary conditions, in particular the estimation of the corresponding bias-terms, will become substantially more complex.

   Nevertheless, results in this direction have been established for specific applications. For example, in [11] and [13], formal proofs have so been esthablished for 'insensitive' product form bounds for finite multi-server tandem queues and queues with overflow respectively. Particulary, in [22] a general framework has been set up to apply the MRA to stochastic service networks under the assumption of continuous service distributions with bounded hazard rates. This framework was used to obtain analytic error bounds as well as ordering results for comparing various (also nón-ordered) $GI|G|c$ queueing systems.

2. **Transient situations.** As the (proof) steps in section 9.3 and following rely upon the recurrent (or dynamic programming) reward relation (9.3.5), similar comparison and error bound results are implicity covered for any finite number of steps, that is finite time horizon of periods each of exponential length with parameter $H = h^{-1}$, hence of average duration h.

   As shown in VDK, by retransforming the time-uniformization (using the Poisson-Gamma relation), in principle these results in turn can also be transformed into comparison and error bound results for any fixed time horizon, say of length T, or to stochastic periods up to exiting or leaving some set of states B (first passage times). Nevertheless, in the latter more practical case, the verification (bounding) step for the bias-terms will generally become harder as different starting states will also have different first passage times.

3. **Non-negative dynamic systems.** In line with the former transient case, as shown in [24], the MRA can also be extended to dynamic systems of the form:

$$\frac{d}{dt}\boldsymbol{W}_t = \boldsymbol{A}\boldsymbol{W}_t$$

where **A** is some arbitrary nonnegative generator rather than a stochastic infinitesimal generator $\boldsymbol{Q}$, such as naturally arising for instance in economic input-output models.

9.7.2 Further Research

1. **A more general verification technique for the bias-terms.** The verification of the bias-terms condition, such as in sections 9.4, 9.5 and 9.6, still appears to be strongly dependent on both the application and its combination with the performance measure (reward function) of interest. A more general ’verification’ technique such as for a class of network configurations or a class of performance measures is still lacking.
2. **Nonexponential queueing networks.** Despite the phase-type approach mentioned and the specific references given above, a simpler and more common extension, such as in line with the sample path comparison approach, to cover non-exponential networks more easily, is sought for.
3. **Discrete-time queueing networks.** Due to digitization the interest in discrete-time rather than continuous-time queueing networks remains growing. Again, in principle the MRA can be set up just as well. However, the appealing property of single moments at a time, as for continuous-time networks might disappear. This will highly complicate the verification of the bias-terms required. In addition, appropriate modifications to guarantee analytic expressions, say of product form type, (such as in [4], [8], [35]) for simplified computations or bounds, will become more difficult to be recognized. In line with practical developments further research in these directions is of substantial interest.
4. **State dependent routing and servicing (call centers/internet).**
   Practical queueing systems can have more complicated state dependent routing or servicing mechanisms than just determined by blocking upon congestion by finite capacity constraints, as used in this chapter.

   As one practical example of interest, in present-day highly developed call centers an incoming call might be routed (so called skill based routing) to the ’best suited (skilled)’ agent group available (by searching through a skill preference list). Conversely, an agent that becomes available might search for the most preferable or suited call waiting.

   A second example that receives considerable interest within present-day queueing (performance evaluation) literature is that of the internet. In this case tandem (packet switch) type structures are used in which service capacity is shared over multiple stations. As a consequence, the service capacity at one station depends on the current loads at other stations. As analytic solutions for these systems are highly limited, the application of the MRA seems of considerable interest. However, due to the state dependent mechanisms the essential step of analytically bounding the corresponding bias-terms will include various complications that are still open.

5. **Mobile server networks (mobile communications, ad-hoc networks, ambulance services).** Another 'class' of unsolvable service (queueing) networks for which the MRA might be useful are service networks in which the users and/or servers themselves are stochastically changing position (moving).

   One example, as already looked at in its simplest form of fixed channels in section 9.6.4, is that of cellular mobile networks in which the users (calling persons) can move during their service (the call) to another service station (a cell) at which they need another service (frequency channel).

   Another present-day example for technical development is that of so-called ad-hoc networks in which transmitters, in a temporarily set up network configuration, highly interact (transmission contentions and loss). In addition, these transmitters may stochastically vary their location (and consequently, interactions).

   As a last example, for modeling ambulance services, with a limited number of servers (the ambulances), both the service times (trip durations), and the availability and the locations (different collection points and hospitals) are subject to stochasticity. A similar remark and appeal for future research of a MRA application also applies here.

9.7.3 Other applications.

All three applications (in sections 9.4, 9.5, 9.6) were based upon the combination of

(i) a modification of an original unsolvable system of practical interest
into a solvable (product form) system, and

(ii) the Markov reward approach to show that this modification leads
to secure bounds or to an error bound for the discrepancy.

This combination has proven to be fruitful in a number of situations. To conclude this chapter below three more applications will be described briefly to illustrate the practical diversity of this combination.

Overflow queues (e.g. Call Centers). In present-day call centers complex skill based routing can be applied. Under skill based routing in its simplest form, incoming calls might be rerouted to a second or higher level agent skill group if a primary access group is not available.

But even in this most simple situation a closed form queueing expression doesn't seem to be available. To be more concrete, consider a standard Erlang loss system with c_1 primary servers without waiting facility ($M|M|c_1|c_1$-queue) in which calls are rerouted to a second finite server group with c_2 (overflow) servers and also no waiting facility ($M|M|c_2|c_2$-queue) if all primary c_1 servers are occupied.

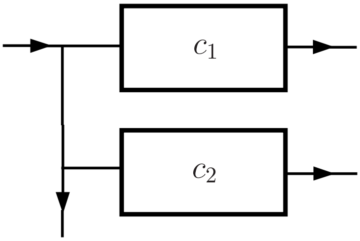


Fig. 9.9: Overflow system.

If also these c_2 servers are busy an incoming call is lost. In addition, assume that a call at one of these c_2 servers remains to be serviced by this server also if meanwhile one of the primary servers becomes idle. In this case, no analytic expression for the loss probability of incoming calls seems to be available. Various approximations have therefore been developed, most notably among them, as based upon the ERM (Equivalent Random Method). However, other than by numerical investigation no accuracy for these approximations has been reported nor can it be secured whether they provide lower or upper bounds.

In [25] (also see chapter 1, section 2.2) a product form modification has therefore been suggested (as by the well-known call packing principle). By the Markov reward approach it was shown that this modification leads to secure (and in fact, quite accurate) upper bounds for the loss probability (as of natural practical interest for dimensioning purposes).

Cellular Mobile Communication Systems. As already described in section 9.6.1, even in its simple form of cells with a fixed number of N_i frequency channels in cell i, due to handover calls from one cell into another, cellular mobile communication system do not exhibit a product form expression (see [5] for a more general and non-exponential description).

In [6] therefore, different product form modifications have been suggested(such as by a redial mechanism). Again, by the Markov reward approach it was shown that these also lead to upper bounds for the various (fresh call and handover) loss probabilities. In this case, however, an intermediate system (other than the original and modified system) had to be used to establish the proofs.

Intensive Care Modeling in Hospitals. The number of intensive care beds is a high cost but also high quality factor for hospitals as it may put lives at risk. This number thus has to be dimensioned carefully. An intensive care bed can be required for either external emergency patients or elective patients that have become critical (±60%) or patients for postoperative care after a ’heavy’ operation (±40%).

A coupling between the operating rooms (operating theater - OT) and the intensive care unit (ICU) is thus intrinsically involved. Due to this coupling, an ICU cannot simply be regarded as an Erlang loss system ($M|M|c|c$-queue) so as to compute the availability (or rejection rate), nor can the OT-ICU be seen as a finite tandem queue as dealt with in section 9.5.

Nevertheless, as outlined in chapter 1, section 8, by the same steps of product form modifications and the Markov reward proof technique (with the technical verification of the bias-terms being quite complicated) in [18] it was shown that the ICU-rejection probability can be approximated reasonably well by an $M|M|c|c$-queue and be bounded from above by an $M|M|c-1|c-1$-queue as of practical interest.

Acknowledgements

Over the years a number of colleagues in the field that I had the pleasure of working with have been most stimulating for continuing on this topic of the Markov Reward Approach. The development of this approach has highly benefitted from their specific motivational applications and comments.

In this respect, with apologies for not being able to be exhaustive, I would like to mention Martin Puterman, Hayri Korezlioglu †, Jan van der Wal, Peter Taylor, Masakiyo Miyazawa and Richard Boucherie.

For his motivation and providing me the roots in both the direction of Markov decision theory and Queueing Networks, I am indebted to Arie Hordijk. Last but not least, I like to acknowledge Nikky Kortbeek for his excellent assistance and collaboration for styling this chapter.

The support by the Dutch Organisation for Scientific Research (NWO) by which this research could be realized has been highly appreciated.

References

1. I. Adan and J. van der Wal (1989a). Monotonicity of the throughput of a closed queueing network in the number of jobs. *Operations Research*, **37**, 953-957.
2. I. Adan and J. van der Wal (1989b). Monotonicity of the throughput in single server production and assembly networks with respect to buffer sizes. *Queueing Networks with Blocking*, eds. H.G. Perros and T. Altiok, North-Holland, 345-356.
3. S. Borst, R.J. Boucherie, O.J. Boxma (1999). ERMR: A generalised equivalent random method for overflow systems with repacking. *ITC* **16**, Elsevier, Amsterdam, eds. P. Key, D. Smith, 313-323.
4. R.J. Boucherie and N.M. van Dijk (1991). Product forms for queueing networks with state-dependent multiple job transitions. *Advances in Applied Probability*, **23**, 152-187.
5. R.J. Boucherie and N.M. van Dijk (2000). On a queueing network for cellular mobile telecommunication networks. *Operations Research*, **48**, 38-49.
6. R.J. Boucherie and N.M. van Dijk (2009). Monotonicity and error bounds for networks of Erlang loss queues. *To appear: Queueing Systems.*
7. S. Brandwajn and J.L. Jow. An approximation method for tandem queues with blocking. *Operations Research*, **36**, 73-83.

8. H. Daduna (2001). *Queueing Networks with Discrete Time Scale*. Springer, Berlin.
9. H. Daduna and R. Szleki (1995). Dependencies in Markovian networks. *Advances in Applied Probability* **25**, 226-254.
10. H. Daduna and R. Szleki (2006). Dependence ordering of Markov Processes. *Journal of Applied Probability*, **43**, 1-22.
11. N.M. van Dijk (1988a). A formal proof for the insensitivity of simple bounds for multiserver non-exponential tandem queues based on monotonicity results. *Stochastic Processes and Applications*, **27**, 261-277.
12. N.M. van Dijk (1988b). Perturbation theory for unbounded Markov reward processes with applications to queueing. *Advances Applied Probability*, **20**, 99-111.
13. N.M. van Dijk (1989). A proof of simple insensitive bounds for a pure overflow system. *Journal of Applied Probability*, **26**, 113-120.
14. N.M. van Dijk (1991). Truncation of Markov chains with applications to queueing. *Operations Research,* **39**, 1018-1026.
15. N.M. van Dijk (1993). *Queueing networks and product forms: a systems approach.* Wiley, Chichester.
16. N.M. Van Dijk (1997). Bounds and error bounds for queueing networks. *Annals of Operations Research*, **79**, 295-319.
17. N.M. van Dijk and K. Korezlioglu (2000). Sensitivity analysis for Markov reward structures until entrance times. *Journal of Applied Probability*, **37**, 45-63.
18. Van Dijk, N.M., Kortbeek, N. (2009). Erlang Loss Bounds for OT-ICU systems. *To appear: Queueing systems*, **67**.
19. N.M. van Dijk and B.F. Lamond (1988). Bounds for the call congestion of finite single-server exponential tandem queues. *Operations Research*, **36**, 470-477.
20. N.M. van Dijk and M. Miyazawa (1997a). A note on bounds and error bounds for nonexponential batch arrival systems. *Probability in the Engineering and Informational Sciences,* **11,** 189-201.
21. N.M. van Dijk and M. Miyazawa (1997b). Error bounds on a practical approximation for finite tandem queues. *Operations Research Letters*, **21**, 201-208.
22. N.M. van Dijk and M. Miyazawa (2004). Error bounds for perturbing nonexponential queues. *Mathematics of Operations Research*, **29**, 525-558.
23. N.M. van Dijk and M.L. Puterman (1988). Perturbation theory for Markov reward processes with applications to queueing systems. *Advances in Applied Probability***20,** (1988), 79-89.
24. N.M. Van Dijk and K. Sladký (200). A note on uniformization for dynamic non-negative systems. *Journal of Applied Probability*, **37**, 329-341.
25. N.M. van Dijk and H.J. van der Sluis (2007). Call Packing Bounds for Overflow Queues. *Performance Evaluation*. To appear.
26. N.M. van Dijk, P. Tsoucas and J. Walrand (1988). Simple bounds and monotonicity of the call congestion of infinite multiserver delay systems, *Probability in the Engineering and Informational Sciences,* **2,** 129-138.
27. N.M. van Dijk and J. van der Wal (1989). Simple bounds and monotonicity results for finite multi-server exponential tandem queues. *Queueing Systems*, **4**, 1-16.
28. M. El-Taha and J.R. Heath (2000). Traffic overflow in loss systems with selective trunk reservation. *Performance Evaluation*, **41**, 295-306.
29. D.E. Everitt and N.W. Macfadyen (1983). Analysis of multicellular mobile radiotelephone systems with loss. *British Telecom. Tech. J.*, **1**, 37-45.
30. S.B. Gershwin (1987). An efficient decomposition method for approximate evaluation of production lines with finite storage space. *Operations Research*, **35**, 291-305.
31. W.K. Grassmann (1991). Finding transient solutions in Markovian event systems through randomization. *Numerical Solutions of Markov Chains*, eds. W.J. Stewart, Marcel Dekker, 357-371.
32. W.K. Grassmann (ed.) (1999). Computational probability. *Kluwer Academic Publishers*.
33. D. Gross and D.R. Miller (1984). The randomisation technique as a modelling tool and solution procedure for transient Markov processes. *Operations Research* , **32**, 343-361.

34. M. Haviv and L. van der Heyden (1984). Perturbation bounds for the stationary probabilities of a finite Markov chain. *Advances in Applied Probability*, **16**, 804-818.
35. W. Henderson and P.G. Taylor (1992). Discrete-time queueing networks with geometric release probabilities. *Advances of Applied Probability*, **24**, 229-233.
36. F.S. Hillier and R.W. Boling (1967). Finite queues in series with exponential or Erlang service times - a numerical approach. *Operations Researh*, **15**, 286-303.
37. A. Hordijk and N.M. van Dijk (1981). Networks of queues with blocking. *Performance*, **81**, ed. K.J. Kylstra, North-Holland, 51-65.
38. A. Hordijk, A. Ridder, Stochastic inequalities for an overflow model. *Journal of Applied Probability*, **24**, 696-708.
39. A. Hordijk, A. Ridder (1988). Insensitivity bounds for the stationary distribution of nonreversible chains. *Journal of Applied Probability*, **25**, 9-20.
40. F.P. Kelly (1979). *Reversibility and Stochastic Networks,* Wiley, Chichester.
41. F.P. Kelly (1991). Loss Networks, *Annals of Applied Probability*, **1**, 319-378.
42. J.G. Kemeny, L. Snell and A.Q. Knapp (1966). *Denumerable Markov chains.* Van Nostrand, Princeton, NJ.
43. W.A. Massey (1987). Stochastic orderings for Markov processes on partially ordered spaces, *Mathematics of Operations Research*, **12**, 350-367.
44. D. McMillan (1991). Traffic modelling and analysis for cellular mobile networks, 13^{th} *International Teletraffic Congress*.
45. C.D. Meyer, Jr. (1980). The condition of a finite Markov chain and perturbation bounds for the limiting probabilities. SIAM, *J. Alg. Disc. Math.*, **1**, 273-283.
46. M. Miyazawa and J.G. Shanthikumar (1991). Monotonicity of the loss probability of single server finite queue with respect to convex order of arrival or service processes. *Probability in the Engineering and Informatical Sciences*, **5**, 43-53.
47. A. Müller and D. Stoyan (2002). Comparison methods for stochastic models and risks, *Wiley*, Chichester.
48. R.O. Onvural and H.C. Perros (1986). On equivalencies of blocking mechanisms in queueing networks with blocking, *Operations Research Letters,* **5**, 293-297.
49. D.L. Pallant and P.G. Taylor (1995). Modeling handovers in cellular mobile networks with dynamical channel allocation, *Operations Research*, **43**, 33-42.
50. M.L. Puterman (1994). *Markov decision processes: discrete dynamic programming.* John Wiley & Son, New York.
51. A. Ridder (1987). *Stochastic inequalities for queues.* Ph.D. Thesis, University of Leiden.
52. S.M. Ross (1970). *Applied probability models with optimization applications*. Holden-Day, San Fransisco.
53. J.G. Shanthikumar and D.D. Yao (1988). Throughput bounds for closed queueing networks with queue-independent service rates. *Performance Evaluation*, **9**, 69-78.
54. P.J. Schweitzer (1968). Perturbation theory and finite Markov chains, *Journal of Applied Probability*, **4**, 401-413.
55. E. Seneta (1968). The principles of truncations in applied probability. *Comm. Math. Univ. Carolina*, **9**, 533-539.
56. E. Seneta (1980). *Non-Negative Matrices and Markov Chains*. Springer, New York.
57. M. Shaked and J.G. Shanthikumar (1994). *Stochastic orders and their applications*. Academic Press, San Diego.
58. D. Sonderman (1979a). Comparing multi-server queues with finite uniting rooms, I: Same number of servers. *Advances in Applied Probability*, **11**, 439-447.
59. D. Sonderman (1979b). Comparing multi-server queues with finite waiting rooms, II: Different number of servers. *Advances in Applied Probability*, **11**, 448-455.
60. W.J. Stewart ed. (1991). *Numerical Solutions of Markov Chains,* . Marcel Dekker.
61. D. Stoyan (1983). *Comparison Methods for Queues and Other Stochastic Models*. Wiley, New York.
62. R. Suri (1985). A concept of monotonicity and its characterization for closed queueing networks. *Operations Researh*, **33**, 606-624.

63. R. Szekli (1995). *Stochastic Ordering and Dependence in Applied Probability*. Lecture Notes in Statist. 97 Springer Verlag, New York.
64. P.G. Taylor and N.M. van Dijk (1998). Strong stochastic bounds for the stationary distribution of a class of multicomponent performability models. *Operations Research*, **46**, 665-674.
65. P. Tsoucas and J. Walrand (1989). Monotonicity of throughput in non-Markovian networks. *Journal of Applied Probability*, **26**, 134-141.
66. J. van der Wal (1989). Monotonicity of the throughput of a closed exponential queueing network in the number of jobs. *OR Spectrum*, **11**, 97-100.
67. W. Whitt (1978). Approximations of dynamic programs. *Mathematics of Operations Research*, **3**, 231-243.
68. W. Whitt (1981). Comparing counting processes and queues. *Advances in Applied Probability*, **13**, 207-220.
69. W. Whitt (1986). Stochastic comparison for non-Markov processes. *Mathemetics of Operations Research*, **11**, 608-618.
70. D.D. Yao (1985). Some properties of the throughput of a closed exponential network of queues. *Operations Research Letters*, **3**, 313-317.

Chapter 10
Stability of Join-the-Shortest-Queue networks: Analysis by Fluid Limits

J. G. Dai, John J. Hasenbein and Bara Kim

Abstract The standard fluid model tool is employed to investigate stability behavior in a variant of a generalized Jackson queueing network. In the network, some customers use a join-the-shortest-queue policy when entering the network or moving to the next station. Furthermore, we allow interarrival and service times to have general distributions. For networks with two stations, necessary and sufficient conditions are given for positive Harris recurrence of the network process. These conditions involve only the mean values of the network primitives. Two counterexamples are provided to show that more information on distributions and tie-breaking probabilities is needed for networks with more than two stations, in order to characterize the stability of such systems. However, if the routing probabilities in the network satisfy a certain homogeneity condition, then it is proved that the stability behavior can be explicitly determined, again using the mean value parameters of the network.

10.1 Join-the-shortest-queue networks

We consider a variant of the classical Jackson queueing network [9, 10]. The main added feature is that an arriving customer may have several routes to choose from

J. G. Dai
H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332
e-mail: dai@gatech.edu

John J. Hasenbein
Graduate Program in Operations Research and Industrial Engineering, Department of Mechanical Engineering, University of Texas at Austin, Austin, Texas, 78712
e-mail: jhas@mail.utexas.edu

Bara Kim
Department of Mathematics, Korea University, Anam-dong, Sungbuk-ku, Seoul, 136-701, Korea
e-mail: bara@korea.ac.kr

R.J. Boucherie and N.M. van Dijk (eds.), *Queueing Networks: A Fundamental Approach*,
International Series in Operations Research & Management Science 154,
DOI 10.1007/978-1-4419-6472-4_10, © Springer Science+Business Media, LLC 2011

at its arrival time. We assume that the customer always chooses to join the shortest queue among a set of allowed queues. In addition, we allow the interarrival and service times to have general distributions, rather than being restricted to the exponential case.

The queueing network model of this chapter is assumed to have $J \geq 1$ stations, with each station consisting of a single server. Each station has a dedicated queue or buffer that holds customers waiting to be served by the station. Let $\mathcal{J} = \{1, \cdots, J\}$ be the set of stations. For each station $i \in \mathcal{J}$, let $\eta_i(n)$ be the service time of the *n*th customer to be served by station *i*. We assume that each station is non-idling, that customers within a buffer are served on a first-in–first-out basis, and that no service is preempted. To describe the external arrival processes, let $\mathcal{P}$ be the class of nonempty subsets of $\mathcal{J}$. For each subset $C \in \mathcal{P}$ of queues, there is an associated exogenous arrival process with interarrival times $\{\xi_C(n) : n \geq 1\}$. We call this arrival process a type-C external arrival process. Upon arriving to the network, each type-C customer joins the shortest queue among all the queues in C, using a tie-breaking rule to be specified shortly. After being served by station i, $i \in \mathcal{J}$, a customer leaves the system with probability $1 - p_i^*$, and becomes a type-C customer with probability p_{iC}, independent of the customer's entire history, where $\sum_{C \in \mathcal{P}} p_{iC} = p_i^*$. When multiple queues are tied for the shortest queue, a tie-breaking rule is needed. We assume that for each subset $B \in \mathcal{P}$ of queues, there is a distribution $\gamma_B = \{\gamma_{B,j} : j \in B\}$. When a customer is to join a shortest queue that is tied by a set B of queues, the customer joins queue j with probability $\gamma_{B,j}$ independently of its history. This type of routing behavior on the part of arriving customers is called Join-the-Shortest-Queue (JSQ) in the literature.

We allow $\xi_C(n) = \infty$ for all n for some C. In this case, the type-C-external arrival process is null. Let

$$\mathcal{E} = \{C \in \mathcal{P} : \text{ the type-}C\text{-external arrival process is non-null}\}.$$

For each $C \in \mathcal{E}$, we assume that $\xi_C = \{\xi_C(n) : n \geq 1\}$ is an independent and identically distributed (i.i.d.) sequence with mean $1/\lambda_C$, and for each station i, $\eta_i = \{\eta_i(n) : n \geq 1\}$ is an i.i.d. sequence with mean $1/\mu_i$. We further assume that the interarrival time sequences, service time sequences, feedback decisions, and tie-breaking decisions are all independent. Additional distributional assumptions on the interarrival times will be specified in Section 10.2. We call λ_C the arrival rates, μ_i the service rates, p_{iC} the feedback probabilities, and $\gamma_{B,j}$ the tie-breaking probabilities of the network. From now on, for purposes of discussion, and stating results, we will refer to the network described above as a *JSQ Network*.

The dynamics of the JSQ network can be described by a continuous time Markov process $X = \{X(t) : t \geq 0\}$, as long as the state space is chosen appropriately. When the interarrival and service time distributions are exponential, $Z = \{(Z_1(t), \ldots, Z_J(t)) : t \geq 0\}$ is such a process, where $Z_i(t)$ is the total number of customers that are either waiting in queue i or being served by station i at time t. This chapter is primarily concerned with the stability of the queueing network. The network is said to be stable if the Markov process X is positive Harris recurrent.

When $\lambda_C = 0$ and $p_{i,C} = 0$ for all $C \in \mathcal{P}$ with more than one element, i.e., customers are never offered a choice of queues to join, the corresponding network is called a generalized Jackson network. Under some minor conditions on the interarrival time distributions, it is known that such a network is stable if and only if the traffic intensity at each station is less than one (see, for example, Dai [2]). The traffic intensity is defined through the first order data of the network, i.e., arrival rates, service rates and feedback probabilities. In particular, the stability of a generalized Jackson network does not depend on the distributions of interarrival and service times. One might hope that for the model of this chapter, the positive Harris recurrence can again be determined by the arrival rates, service rates and the feedback probabilities. Theorem 10.3.1 in Section 10.3 shows that this assertion is indeed true when $J = 2$ by describing explicit recurrence conditions in terms of arrival rates, service rates and feedback probabilities. In particular, the stability of a 2-station network does not depend on the distributions of interarrival and service times, nor does it depend on the tie-breaking probabilities. Unfortunately, when $J \geq 3$, an analogous result does not hold. Specifically, two counterexamples in Section 10.4 demonstrate that Theorem 10.3.1 cannot be generalized to larger networks. In the first example with $J = 3$, we show that the positive Harris recurrence of the process depends on the tie-breaking probabilities $\gamma_{B,i}$. In the second example, again with $J = 3$, we show that the positive Harris recurrence of the process depends on the distributions of the service times. However if all the stations have homogeneous feedback probabilities, i.e., if p_{iC} does not depend on queue i, the positive Harris recurrence is again determined by the arrival rates, services and feedback probabilities, and not on the distributions of the interarrival and service times or the tie-breaking probabilities. In this case, Theorem 10.5.2 in Section 10.5 gives explicit recurrence conditions in terms of arrival rates, service rates and feedback probabilities.

Queueing systems with JSQ type routing have a long history in the literature. We only mention the papers in which there is stability analysis of JSQ networks. Kurkova [11] treated a special system when $J = 2$, the interarrival and service times distributions are exponential, and a fair coin is flipped to break a tie. She represented the system as a continuous time Markov chain with a countable state space and obtained an explicit recurrence condition for the Markov chain by using Lyapunov functions. Stability of JSQ networks, when there is no feedback, was studied by Foss and Chernova [8] and Foley and McDonald [7]. A quite general JSQ network with feedback was treated by Suhov and Vvedenskaya [14]. However, their stability analysis was limited to a few special cases.

Queueing networks with alternate routes arise in many telecommunication and service systems. A customer call center is an example of such a service system. The myopic join-the-shortest-queue routing decision is often employed in practice. The stability of these networks is essential to the capacity planning of these systems.

We employ the standard fluid model tool in our stability analysis. Whenever appropriate, we do not go through every detail of using the tool; readers may consult, for example, Dai [4] for additional details. Fluid models are commonly used to prove the positive recurrence of queueing networks and/or the transience of such systems, but here we are also able to use the fluid model approach to prove non-positive recurrence. As such we are able to identify the stability behavior on the boundary of the stability region. The behavior on the boundary is often left as an open question in stability analysis via fluid models, and the method employed in this chapter is quite a new technique to use fluid models to prove the non-positive recurrence of a queueing network.

The chapter is organized as follows: In Section 10.2, we first provide the Markov process characterization of the network. A fluid model for the system is then defined and criteria for stability and instability of the system are given. Section 10.3 gives the necessary and sufficient conditions for stability in terms of arrival rates, service rates and feedback probabilities for systems with two stations ($J = 2$). In Section 10.4, two examples with three stations ($J = 3$) are given, the first of which shows that the stability depends on the tie-breaking probabilities, and the second of which shows that the stability depends on not only the first order data but also the distributions of the service times. In Section 10.5, for systems with more than two stations ($J \geq 3$) that satisfies an additional assumption that all stations have homogeneous feedback probability, we give the necessary and sufficient conditions for stability in terms of arrival rates, service rates and feedback probabilities for systems with two stations Further study on the stability of JSQ networks is described in Section 10.6.

Now we collect some mathematical notation used in the rest of the chapter. For a set C, $|C|$ indicates the cardinality of C. However, for $x \in \mathbb{R}^N$, we use $|x|$ to denote the l^1-norm. For random variables X and Y, $X \geq_{\mathrm{st}} Y$ indicates that X is stochastically larger than Y. When a probability operator appears with a subscript π, this indicates the probability is the one generated by initial distribution π (this may include a degenerate initial distribution consisting of only one state).

10.2 The network process and the fluid model

We use

$$X(t) = (Z(t), U(t), V(t)) \tag{10.1}$$

to denote the state of our queueing network at time t. The first component $Z(t) = (Z_1(t), \cdots, Z_J(t))$ is J-dimensional, where, as before, $Z_i(t)$ is the total number of customers that are either waiting in queue i or being served by station i at time t. The second component $U(t) = (U_C(t) : C \in \mathcal{E})$ is $|\mathcal{E}|$-dimensional, where $U_C(t)$ is the remaining interarrival time of the type-C external arrival process at time t. The last component $V(t) = (V_1(t), \cdots, V_J(t))$ is J-dimensional, where $V_i(t)$ is the remaining service time of the customer who is in service at station i at time t. ($V_i(t)$ is set to be zero if there is no customer in service at station i at time t.) The process $X = \{X(t) : t \geq 0\}$ is taken to be right continuous with left limits. It follows from Dai [2] that X is a strong Markov process whose state space $\mathcal{S}$ is a subset of $\mathbb{R}^{2J+|\mathcal{E}|}$.

The Markov process X is said to be positive Harris recurrent if it possesses a unique stationary distribution. To apply the fluid limit technique to the stability analysis, we make the following additional assumptions on interarrival times. We assume that, for any $C \in \mathcal{E}$, the distribution of $\xi_C(1)$ is unbounded, i.e.,

$$\mathbb{P}(\xi_C(1) \geq x) > 0, \text{ for any } x > 0. \tag{10.2}$$

We also assume that, for any $C \in \mathcal{E}$, the distribution of $\xi_C(1)$ is spread out, i.e., there exists an integer $n_C > 0$ and a function $q_C(x) \geq 0$ on $(0,\infty)$ with $\int_0^\infty q_C(x)dx > 0$, such that

$$\mathbb{P}(a \leq \xi_C(1) + \cdots + \xi_C(n_C) \leq b) \geq \int_a^b q_C(x)dx, \text{ for any } 0 \leq a < b.$$

We now introduce the queueing and fluid dynamical equations, and provide results which relate the queueing model and fluid models defined by these equations. This framework allows us to use fluid model techniques to prove the results on stability of the JSQ networks in subsequent sections.

We define a number of processes related to the queueing network:

$E(t) = \{E_C(t) : C \in \mathcal{E}\}, t \geq 0$, where $E_C(t)$ is the number of customers which arrive during $[0,t]$ due to the type-C external arrival process.

$A(t) = \{A_i(t) : i \in \mathcal{J}\}, t \geq 0$, where $A_i(t)$ is the number of arrivals to buffer i during $[0,t]$ (including exogenous arrivals and feedback arrivals).

$D(t) = \{D_i(t) : i \in \mathcal{J}\}, t \geq 0$, where $D_i(t)$ is the number of customers which complete service at station i during $[0,t]$.

$S(t) = \{S_i(t) : i \in \mathcal{J}\}, t \geq 0$, where $S_i(t)$ is the number of customers station i completes if it spends t units of time working on such customers.

$\Phi(n)\{\Phi_{iC}(n) : i \in \mathcal{J}, C \in \mathcal{P}\}, n = 0,1,2,\cdots$, where $\Phi_{iC}(n)$ is the number of customers, among the first n who depart station i, which become type-C customers.

$T(t) = \{T_i(t) : i \in \mathcal{J}\}, t \geq 0$, where $T_i(t)$ is the amount of time spent working on customers at station i during $[0,t]$.

$I(t) = \{I_i(t) : i \in \mathcal{J}\}, t \geq 0$, where $I_i(t)$ is the amount of time station i idles during $[0,t]$.

Then, the following equations define the dynamics of a JSQ network: For $i \in \mathcal{J}$ and $0 \leq s \leq t$,

$$Z_i(t) = Z_i(0) + A_i(t) - D_i(t), \tag{10.3}$$

$$Z_i(t) \geq 0, \tag{10.4}$$

$$T_i(\cdot) \text{ and } I_i(\cdot) \text{ are nondecreasing}, \tag{10.5}$$

$$T_i(t) + I_i(t) = t, \tag{10.6}$$

$$\text{If } Z_i(u) > 0 \text{ for } u \in (s,t), \text{ then } I_i(s) = I_i(t). \tag{10.7}$$

$$D_i(t) = S_i(T_i(t)). \tag{10.8}$$

For $C \in \mathcal{P}$ and $0 \le s \le t$,

$$\begin{aligned}&\sum_{i\in C}(A_i(t) - A_i(s))\\ &\quad \ge \sum_{B:B\subseteq C}\left\{(E_B(t) - E_B(s)) + \sum_{i\in\mathcal{J}}(\Phi_{iB}(D_i(t)) - \Phi_{iB}(D_i(s)))\right\}.\end{aligned} \tag{10.9}$$

For $C \in \mathcal{P}$ and $0 \le s \le t$, if $Z_i(u) > Z_j(u)$ for all $i \in C$, $j \in \mathcal{J} - C$ and $u \in (s,t)$, then

$$\begin{aligned}&\sum_{i\in C}(A_i(t) - A_i(s))\\ &\quad = \sum_{B:B\subseteq C}\left\{(E_B(t) - E_B(s)) + \sum_{i\in\mathcal{J}}(\Phi_{iB}(D_i(t)) - \Phi_{iB}(D_i(s)))\right\}.\end{aligned} \tag{10.10}$$

Equations (10.3)-(10.8) are standard equations for generalized Jackson networks operating under an arbitrary non-idling policy. The last two equations however, are new, and they enforce the JSQ routing behavior of the customers.

Using the dynamical equations (10.3)-(10.10) we derive the corresponding fluid model equations. Our methodology closely follows a now standard procedure and we only outline the general steps. By the strong law of large numbers, for almost all sample paths ω, we have

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\xi_C(k,\omega) = \lambda_C^{-1}, \quad C \in \mathcal{E}, \tag{10.11}$$

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\eta_i(k,\omega) = \mu_i^{-1}, \quad i \in \mathcal{J}, \tag{10.12}$$

$$\lim_{n\to\infty}\frac{1}{n}\Phi_{iC}(n,\omega) = p_{iC}, \quad i \in \mathcal{J}, C \in \mathcal{P}. \tag{10.13}$$

Let $\mathbb{X} \equiv \{(A(t),T(t),I(t),Z(t)), t \ge 0\}$ be a network process governed by (10.3)-(10.10), and $\mathbb{X}_x$ be such a process with initial state $x = (z,u,v)$. By taking $C = \mathcal{J}$ in (10.10), one has for each station k and each $0 \le s < t$ that

$$\begin{aligned}&A_k(t) - A_k(s)\\ &\quad \le \sum_{i\in\mathcal{J}}(A_i(t) - A_i(s))\\ &\quad = \sum_{B:B\subseteq\mathcal{J}}\left\{(E_B(t) - E_B(s)) + \sum_{i\in\mathcal{J}}(\Phi_{iB}(D_i(t)) - \Phi_{iB}(D_i(s)))\right\}.\end{aligned}$$

It follows from the same argument as in Dai [2] that for every sample path ω satisfying (10.11)-(10.13) and every collection $\{x_r : r > 0\}$ of initial states such that $\{|x_r|/r : r > 0\}$ is bounded, there exists a subsequence $r_n \to \infty$ such that $\frac{1}{r_n}\mathbb{X}_{x_{r_n}}(r_n\cdot,\omega)$ converges uniformly on any compact subset of $[0,\infty)$ to some limit say $\bar{\mathbb{X}} = (\bar{A}(\cdot),\bar{T}(\cdot),\bar{I}(\cdot),\bar{Z}(\cdot))$. Each such limit $\bar{\mathbb{X}}$ is called a *fluid limit*. In the special case where the sequence of initial states $\{x_r : r > 0\}$ is independent of r, we call the limit a *fluid limit with fixed initial state*. Both types of fluid limits are used in our subsequent stability analysis of the process $\mathbb{X}$.

As shown in Bramson [1], in the analysis of stability via fluid limits, it is sufficient to consider the so-called undelayed fluid limit, i.e. when

$$\lim_{r\to\infty} \frac{1}{r}(|u_r| + |v_r|) = 0, \tag{10.14}$$

where u_r and v_r are subvectors of the initial state $x_r = (z_r, u_r, v_r)$. Thus, from now on we only consider undelayed fluid limits.

Now, let $\bar{\mathbb{X}}$ be a fluid limit obtained from a sequence of initial states $\{x_r\}$ satisfying (10.14). It is readily seen that all of $\bar{A}_i(\cdot), \bar{T}_i(\cdot), \bar{I}_i(\cdot)$ and $\bar{Z}_i(\cdot)$, $i \in \mathcal{J}$, are Lipschitz continuous. Hence they are absolutely continuous and thus differentiable almost everywhere with respect to the Lebesgue measure. We say that t is a regular point of $\bar{\mathbb{X}}$ if all components of $\bar{\mathbb{X}}$ are differentiable at t. From now on, we implicitly assume that t is a regular point whenever the derivative of a component of $\bar{\mathbb{X}}$ is involved. Applying fluid limits to (10.3)-(10.14), we obtain the equations: For $i \in \mathcal{J}$ and $t \geq 0$,

$$\bar{Z}_i(t) = \bar{Z}_i(0) + \bar{A}_i(t) - \mu_i \bar{T}_i(t), \tag{10.15}$$

$$\bar{Z}_i(t) \geq 0, \tag{10.16}$$

$$\bar{T}_i(\cdot) \text{ and } \bar{I}_i(\cdot) \text{ are nondecreasing}, \tag{10.17}$$

$$\bar{T}_i(t) + \bar{I}_i(t) = t, \tag{10.18}$$

$$\text{If } \bar{Z}_i(t) > 0, \text{ then } \dot{\bar{I}}_i(t) = 0. \tag{10.19}$$

For $C \in \mathcal{P}$ and $t \geq 0$,

$$\sum_{i\in C} \dot{\bar{A}}_i(t) \geq \Lambda_C + \sum_{i\in\mathcal{J}} P_{iC}\mu_i \dot{\bar{T}}_i(t). \tag{10.20}$$

For $C \in \mathcal{P}$ and $t \geq 0$, if $Z_i(t) > Z_j(t)$ for all $i \in C$ and $j \in \mathcal{J} \setminus C$, then

$$\sum_{i\in C} \dot{\bar{A}}_i(t) = \Lambda_C + \sum_{i\in\mathcal{J}} P_{iC}\mu_i \dot{\bar{T}}_i(t). \tag{10.21}$$

where

$$\Lambda_C \equiv \sum_{B:\phi\neq B\subset C} \lambda_B \text{ and } P_{iC} \equiv \sum_{B:\phi\neq B\subset C} p_{iB}.$$

We call the equations (10.15)-(10.21) the *fluid model equations* and call a solution $\bar{\mathbb{X}} = \{(\bar{A}(t), \bar{T}(t), \bar{I}(t), \bar{Z}(t)), t \geq 0\}$, of the fluid model equations a *fluid model solution*. Note that any fluid limit with fixed initial state necessarily has $\bar{Z}(0) = 0$. Thus these fluid limits form a subset of fluid solutions with $\bar{Z}(0) = 0$. The following definitions and lemmas indicate the usefulness of different types of fluid limits.

Definition 10.2.1 *(i) The fluid model is stable if there exists a $\delta > 0$ such that for each fluid model solution $\bar{\mathbb{X}}$, with $|\bar{Z}(0)| \leq 1$, $\bar{Z}(t) = 0$ for $t \geq \delta$.*
(ii) The fluid model is weakly unstable if there exists a $\delta > 0$ such that for each fluid model solution $\bar{\mathbb{X}}$, with $\bar{Z}(0) = 0$, $\bar{Z}(\delta) \neq 0$.

The same reasoning used in Dai [2, 3], can be applied to the class of networks we consider here to give the following criteria.

Lemma 10.2.2 *(Dai [2]) If the fluid model is stable, then the Markov process X is positive Harris recurrent.*

Lemma 10.2.3 *(Dai [3]) If the fluid model is weakly unstable, then the process X is unstable in the sense that, for each fixed initial state x, $|Z(t)| \to \infty$ as $t \to \infty$ with probability 1.*

If we assume *a priori* that the process X is positive recurrent, then any fluid limit with fixed initial state must obey an extra dynamical equation, which augments the fluid model equations presented in (10.15)-(10.21). It turns out that the augmented set of equations will be quite useful for proving non-positive recurrence using fluid model analysis.

So, suppose that X is positive Harris recurrent and let π be its stationary distribution. Since every station is nonidling, for each fixed initial state x,

$$\begin{aligned}\lim_{t\to\infty} \frac{T_i(t)}{t} &= \lim_{t\to\infty} \frac{1}{t} \int_0^t 1_{\{Z_i(s)>0\}} ds \\ &= \pi\left(\{(z,u,v) \in \mathcal{S} : z_i > 0\}\right) \quad \mathbb{P}_x\text{-a.s.}, i \in \mathcal{J}.\end{aligned}$$

Therefore,

$$\bar{T}_i(t) = t\, \pi\left(\{(z,u,v) \in \mathcal{S} : z_i > 0\}\right), \quad t \geq 0, i \in \mathcal{J}, \tag{10.22}$$

for every fluid limit $\bar{\mathbb{X}}$, which is a limit of scaled sample paths with a fixed initial state. Choose a compact set $\mathcal{K} \subset \mathcal{S}$ such that $\pi(\mathcal{K}) > 0$. By (10.2), there exists a $t_0 > 0$ such that for each $(z,u,v) \in \mathcal{K}$,

$$\mathbb{P}_{(z,u,v)}(|Z(t_0)| = 0) > 0.$$

Therefore

$$\begin{aligned}\pi(\{(z,u,v) \in \mathcal{S} : |z| = 0\}) &= \mathbb{P}_\pi(|Z(0)| = 0) \\ &= \mathbb{P}_\pi(|Z(t_0)| = 0) \\ &= \int_{\mathcal{S}} \mathbb{P}_{(z,u,v)}(|Z(t_0)| = 0)\, d\pi(z,u,v) \\ &\geq \int_{\mathcal{K}} \mathbb{P}_{(z,u,v)}(|Z(t_0)| = 0)\, d\pi(z,u,v) \\ &> 0.\end{aligned}$$

Combining this with (10.22) yields

$$\dot{\bar{T}}_i(t) < 1, \;\; i \in \mathfrak{J}, \tag{10.23}$$

for every fluid limit $\bar{\mathbb{X}}$, which is a limit of scaled sample paths with fixed initial state.

We call the equations (10.15)-(10.21) plus (10.23) the *augmented fluid model equations* and call a solution $\bar{\mathbb{X}}$, to these equations an *augmented fluid model solution*.

Definition 10.2.4 *The augmented fluid model is weakly unstable if there exists a $\delta > 0$ such that for each augmented fluid model solution $\bar{\mathbb{X}}$, with $\bar{Z}(0) = 0$, $\bar{Z}(\delta) \neq 0$.*

Suppose that the augmented fluid model is weakly unstable but the Markov process X is positive Harris recurrent. Since the augmented fluid model equations are satisfied by every fluid limit which is a limit of scaled sample paths with fixed initial state, the argument in Dai [3] implies that the process is unstable in the sense that, $|Z(t)| \to \infty$ as $t \to \infty$ with probability 1, which is a contradiction. Therefore we obtain the following instability criterion.

Lemma 10.2.5 *If the augmented fluid model is weakly unstable, then the Markov process $\{X(t) : t \geq 0\}$ is not positive Harris recurrent.*

10.3 JSQ networks with two stations

For simplicity of notation we use λ_1, λ_2 and λ instead of $\lambda_{\{1\}}$, $\lambda_{\{2\}}$ and $\lambda_{\{1,2\}}$, respectively, in the two station case. We also use p_{ij} instead of $p_{i\{j\}}$, $i, j = 1, 2$. To avoid trivial cases, we assume that $p_{11} < 1$, $p_{22} < 1$ and at least one of p_1^* and p_2^* is less than 1. However, we make no assumptions on $p_{i,\{1,2\}}$, $i = 1, 2$.

The following theorem provides a necessary and sufficient condition for the Markov process X to be positive Harris recurrent.

Theorem 10.3.1 *Consider a JSQ network with $J = 2$. The Markov process X is positive Harris recurrent if and only if the following three conditions hold:*

(i) $\lambda_1 + \lambda_2 + \lambda + (p_1^* - 1)\mu_1 + (p_2^* - 1)\mu_2 < 0$*;*

(ii) if $p_2^* < 1$*, then*

$$p_{21}(\lambda_1 + \lambda_2 + \lambda + \mu_1 p_1^* - \mu_1) + (1 - p_2^*)(\lambda_1 + \mu_1 p_{11} - \mu_1) < 0;$$

(iii) if $p_1^* < 1$*, then*

$$p_{12}(\lambda_1 + \lambda_2 + \lambda + \mu_2 p_2^* - \mu_2) + (1 - p_1^*)(\lambda_2 + \mu_2 p_{22} - \mu_2) < 0.$$

Kurkova [11] obtained a necessary and sufficient condition that is equivalent to ours (see Appendix 2 of Dai, Hasenbein and Kim [6]). Her paper examines the special case when the exogenous arrival processes are Poisson, all service times have an exponential distribution with mean 1, and $\gamma_{\{1,2\},j} = \frac{1}{2}, j = 1, 2$.

Before we prove Theorem 10.3.1, we provide an interpretation of the conditions in Theorem 10.3.1. The first condition is the most straightforward. First, partition the state space of the number of customers in the system, say (z_1,z_2) into two regions. Let region I be $\{(z_1,z_2) : z_1 < z_2\}$ and region II be $\{(z_1,z_2) : z_1 > z_2\}$ (we ignore the boundary set for now). In region I all type-$\{1,2\}$ customers join the queue at station 1. Then, if station 1 is busy the net rate at which it eliminates jobs from the system is

$$r_1 \equiv \mu_1 + \mu_1 p_{12} - \mu_1 p_1^* + \mu_2 p_{22} - \mu_2 p_2^* - \lambda_1 - \lambda .$$

Similarly, in region I the net rate at which station 2 eliminates customers from the system is

$$r_2 \equiv \mu_2 - \mu_1 p_{12} - \mu_2 p_{22} - \lambda_2 .$$

Notice that the left-hand side of condition (i) is simply $-(r_1+r_2)$, i.e. condition (i) implies that the total net rate at which customers are eliminated must be positive. One can check that the left-hand side of (i) also corresponds to the net customer elimination rate in region II. On the boundary between the two regions, the elimination rate seemingly should depend on the tie-breaking probability. However, since the rates are the same in either region, we see that the tie-breaking probability is immaterial to this rate condition.

Condition (i) is a type of drift condition on the interior of the state space. The other two conditions are drift rate conditions on the boundaries. To see this suppose $z_1 = 0$, i.e. station 1 is idle. In this case, the net drift rate of the number of jobs is given by

$$(s_1,s_2) \equiv (\lambda_1 + \lambda + \mu_2(p_2^* - p_{22}), \lambda_2 + \mu_2 p_{22} - \mu_2).$$

Then condition (iii) is equivalent to $(-r_2,r_1)\cdot(s_1,s_2) < 0$, i.e. the normal to the interior drift and the reflection vector must form an acute angle. This is the usual stability condition for a process with (constant) oblique reflection at the boundaries. Condition (ii) has an analogous interpretation for the boundary defined by $z_2 = 0$.

Proof of Theorem 10.3.1.

Sufficiency: Suppose that $\bar{\mathbb{X}}$ is a fluid model solution. Let $f(t) = |\bar{Z}(t)|$. It is readily seen that the fluid model is stable if there exists an $\varepsilon > 0$ such that

$$\dot{f}(t) \le -\varepsilon \ \text{ if } f(t) > 0. \tag{10.1}$$

Hence, by Lemma 10.2.2, X is positive Harris recurrent if there exists an $\varepsilon > 0$ satisfying (10.1). By (10.15), $\dot{f}(t)$ can be written as

$$\dot{f}(t) = \dot{\bar{A}}_1(t) + \dot{\bar{A}}_2(t) - \mu_1 \dot{\bar{T}}_1(t) - \mu_2 \dot{\bar{T}}_2(t).$$

Employing (10.21) with $C = \{1,2\}$ we obtain,

$$\dot{f}(t) = \lambda_1 + \lambda_2 + \lambda + (p_1^* - 1)\mu_1 \dot{\bar{T}}_1(t) + (p_2^* - 1)\mu_2 \dot{\bar{T}}_2(t). \tag{10.2}$$

Now we show that (10.1) holds for some $\varepsilon > 0$ by considering three cases separately.

Case 1. Suppose $\bar{Z}_1(t) > 0$ and $\bar{Z}_2(t) > 0$. Then by (10.18) and (10.19), $\dot{\bar{T}}_i = 1$, $i = 1,2$. So (10.2) becomes

$$\dot{f}(t) = \lambda_1 + \lambda_2 + \lambda + (p_1^* - 1)\mu_1 + (p_2^* - 1)\mu_2,$$

which is negative by (i).

Case 2. $\bar{Z}_1(t) > 0$ and $\bar{Z}_2(t) = 0$.

By (10.18) and (10.19),

$$\dot{\bar{T}}_1(t) = 1. \tag{10.3}$$

Substituting (10.3) into (10.2) gives,

$$\dot{f}(t) = \lambda_1 + \lambda_2 + \lambda + (p_1^* - 1)\mu_1 + (p_2^* - 1)\mu_2 \dot{\bar{T}}_2(t). \tag{10.4}$$

Next, evaluating (10.21) with $C = \{1,2\}$ and using (10.3) yields

$$\dot{\bar{A}}_1(t) + \dot{\bar{A}}_2(t) = \lambda_1 + \lambda_2 + \lambda + p_1^*\mu_1 + p_2^*\mu_2\dot{\bar{T}}_2(t). \tag{10.5}$$

Similarly, evaluating (10.21) with $C = \{1\}$, along with (10.3) yields

$$\dot{\bar{A}}_1(t) = \lambda_1 + p_{11}\mu_1 + p_{21}\mu_2\dot{\bar{T}}_2(t). \tag{10.6}$$

We subtract (10.6) from (10.5) to obtain

$$\dot{\bar{A}}_2(t) = \lambda_2 + \lambda + (p_1^* - p_{11})\mu_1 + (p_2^* - p_{21})\mu_2\dot{\bar{T}}_2(t). \tag{10.7}$$

By assumption $\bar{Z}_2(t) = 0$ which implies $\dot{\bar{Z}}_2(t) = 0$. Hence by (10.15),

$$\dot{\bar{A}}_2(t) = \mu_2\dot{\bar{T}}_2(t). \tag{10.8}$$

Therefore, substituting (10.8) into (10.7) gives

$$\mu_2(1 - p_2^* + p_{21})\dot{\bar{T}}_2(t) = \lambda_2 + \lambda + (p_1^* - p_{11})\mu_1. \tag{10.9}$$

Now, if $1 - p_2^* + p_{21} = 0$ then $p_2^* = 1$ and so, by (10.4),

$$\dot{f}(t) = \lambda_1 + \lambda_2 + \lambda + (p_1^* - 1)\mu_1,$$

which is negative by (i).

Otherwise, suppose $1 - p_2^* + p_{21} > 0$. Then, by (10.9),

$$\dot{\bar{T}}_2(t) = \frac{\lambda_2 + \lambda + (p_1^* - p_{11})\mu_1}{\mu_2(1 - p_2^* + p_{21})}. \tag{10.10}$$

In this case, (10.10) and (10.4) imply

$$\dot{f}(t) = \frac{p_{21}[\lambda_1+\lambda_2+\lambda+\mu_1(p_1^*-1)]+(1-p_2^*)[\lambda_1+\mu_1(p_{11}-1)]}{1-p_2^*+p_{21}},$$

which by (i) is negative when $p_2^*=1$ and by (ii) is negative when $p_2^*<1$.

Case 3. $\bar{Z}_1(t)=0$ and $\bar{Z}_2(t)>0$. The argument in this case is analogous to that of case 2.

Necessity: Lemma 10.2.5 implies that we need only show that the augmented fluid model is weakly unstable if any of (i)-(iii) of Theorem 10.3.1 does not hold. By symmetry, it is sufficient to analyze the three cases examined below. Let $\bar{\mathbb{X}}$ be an augmented fluid model solution with $\bar{Z}(0)=0$ and let

$$f(t) = |\bar{Z}(t)|, \ \ t \geq 0.$$

Considering three cases separately, we show that $\dot{f}(t)>0$ for all regular $t>0$, which completes the proof.

Case 1. Suppose (i) does not hold. By (10.2) and (10.23),

$$\dot{f}(t) > \lambda_1+\lambda_2+\lambda+(p_1^*-1)\mu_1+(p_2^*-1)\mu_2 \ \geq \ 0,$$

which proves the result for this case.

Case 2. Suppose (i) holds and (ii) does not hold. If $\bar{Z}_2(t)>0$, then by (10.18) and (10.19), $\dot{\bar{T}}_2(t)=1$, which contradicts (10.23). Hence $\bar{Z}_2(t)=0$ and $\dot{\bar{Z}}_2(t)=0$. As before, by (10.15),

$$\dot{\bar{A}}_2(t) = \mu_2\dot{\bar{T}}_2(t). \tag{10.11}$$

By subtracting (10.20) evaluated at $C=\{1\}$ from (10.21) evaluated at $C=\{1,2\}$, we have

$$\dot{\bar{A}}_2(t) \leq \lambda_2+\lambda+(p_1^*-p_{11})\mu_1\dot{\bar{T}}_1(t)+(p_2^*-p_{21})\mu_2\dot{\bar{T}}_2(t).$$

Hence by (10.23) and (10.11),

$$\dot{\bar{T}}_2(t) < \frac{\lambda_2+\lambda+(p_1^*-p_{11})\mu_1}{\mu_2(1-p_2^*+p_{21})}. \tag{10.12}$$

Substituting (10.12) into (10.2) and applying $\dot{\bar{T}}_1(t)<1$ lead to

$$\dot{f}(t) > \frac{p_{21}(\lambda_1+\lambda_2+\lambda+\mu_1p_1^*-\mu_1)+(1-p_2^*)(\lambda_1+\mu_1p_{11}-\mu_1)}{1-p_2^*+p_{21}}.$$

The numerator above is nonnegative by the negation of (ii), thus $\dot{f}(t)>0$. □

The following theorem provides a sufficient condition for the Markov process X to be unstable in the sense that $|Z(t)| \to \infty$ as $t \to \infty$ with probability 1.

Theorem 10.3.2 *Consider a JSQ network with $J = 2$. The process X is unstable in the sense that $|Z(t)| \to \infty$ as $t \to \infty$ with probability 1 if*

$$\lambda_1 + \lambda_2 + \lambda + (p_1^* - 1)\mu_1 + (p_2^* - 1)\mu_2 > 0, \tag{10.13}$$
$$or \;\; p_{21}(\lambda_1 + \lambda_2 + \lambda + \mu_1 p_1^* - \mu_1) + (1 - p_2^*)(\lambda_1 + \mu_1 p_{11} - \mu_1) > 0, \tag{10.14}$$
$$or \;\; p_{12}(\lambda_1 + \lambda_2 + \lambda + \mu_2 p_2^* - \mu_2) + (1 - p_1^*)(\lambda_2 + \mu_2 p_{22} - \mu_2) > 0. \tag{10.15}$$

Proof. Suppose that $\bar{\mathbb{X}}$ is a fluid model solution with $\bar{Z}(0) = 0$, $t \ge 0$. Let $f(t) = |\bar{Z}(t)|$. By Lemma 10.2.3, it suffices to show that $\dot{f}(t) > 0$ for all $t > 0$. We show this by considering three cases separately.

Case 1. Suppose (10.13) holds.
Since $\dot{\bar{T}}_1(t) \le 1$ and $\dot{\bar{T}}_2(t) \le 1$, by (10.2),

$$\dot{f}(t) \ge \lambda_1 + \lambda_2 + \lambda + (p_1^* - 1)\mu_1 + (p_2^* - 1)\mu_2 > 0,$$

for all $t > 0$.

Case 2. Suppose (10.13) does not hold and (10.14) holds.
If $p_2^* = 1$, then $\dot{f}(t) \ge \lambda_1 + \lambda_2 + \lambda + (p_1^* - 1)\mu_1 > 0$ by (10.2) and (10.14). Now suppose that $p_2^* < 1$. First we show that

$$\bar{Z}_2(t) = 0, \;\; t \ge 0. \tag{10.16}$$

To prove (10.16), it suffices to show that $\dot{\bar{Z}}_2(t) \le 0$ if $\bar{Z}_2(t) > 0$. Suppose $\bar{Z}_2(t) > 0$. Then by (10.18) and (10.19), $\dot{\bar{T}}_2(t) = 1$. By (10.15),

$$\dot{\bar{A}}_2(t) = \dot{\bar{Z}}_2(t) + \mu_2. \tag{10.17}$$

By subtracting (10.20) evaluated at $C = \{1\}$ from (10.21) evaluated at $C = \{1,2\}$, we have

$$\dot{\bar{A}}_2(t) \le \lambda_2 + \lambda + (p_1^* - p_{11})\mu_1 \dot{\bar{T}}_1(t) + (p_2^* - p_{21})\mu_2. \tag{10.18}$$

Substituting (10.17) into (10.18) and applying $\dot{\bar{T}}_1(t) \le 1$ lead to

$$\dot{\bar{Z}}_2(t) \le \lambda_2 + \lambda + (p_1^* - p_{11})\mu_1 + (p_2^* - p_{21} - 1)\mu_2. \tag{10.19}$$

Since by assumption,(10.13) does not hold,

$$\mu_2 \ge \frac{\lambda_1 + \lambda_2 + \lambda + \mu_1(p_1^* - 1)}{1 - p_2^*}. \tag{10.20}$$

Finally, by (10.19) and (10.20),

$$\dot{\bar{Z}}_2(t) \le -\frac{p_{21}(\lambda_1+\lambda_2+\lambda+\mu_1 p_1^* - \mu_1) + (1-p_2^*)(\lambda_1+\mu_1 p_{11}-\mu_1)}{1-p_2^*}.$$

Hence, $\dot{\bar{Z}}_2(t) < 0$ by (10.14). Thus (10.16) holds.

Next, subtracting (10.20) evaluated at $C=\{1\}$ from (10.21) evaluated at $C=\{1,2\}$, we obtain

$$\begin{aligned}\dot{\bar{A}}_2(t) &\le \lambda_2+\lambda+(p_1^*-p_{11})\mu_1\dot{\bar{T}}_1(t)+(p_2^*-p_{21})\mu_2\dot{\bar{T}}_2(t)\\ &\le \lambda_2+\lambda+(p_1^*-p_{11})\mu_1+(p_2^*-p_{21})\mu_2\dot{\bar{T}}_2(t). \qquad (10.21)\end{aligned}$$

By (10.16), $\dot{\bar{Z}}_2(t)=0$ and so $\dot{\bar{A}}_2(t)=\mu_2\dot{\bar{T}}_2(t)$ by (10.15). Hence, employing (10.21), we have

$$\dot{\bar{T}}_2(t) \le \frac{\lambda_2+\lambda+(p_1^*-p_{11})\mu_1}{\mu_2(1-p_2^*+p_{21})}. \qquad (10.22)$$

Substituting (10.22) into (10.2) and applying $\dot{\bar{T}}_1(t)\le 1$ lead to

$$\dot{f}(t) \ge \frac{p_{21}(\lambda_1+\lambda_2+\lambda+\mu_1 p_1^* - \mu_1) + (1-p_2^*)(\lambda_1+\mu_1 p_{11}-\mu_1)}{1-p_2^*+p_{21}}.$$

Thus (10.14) now implies $\dot{f}(t)>0$.

Case 3. Suppose (10.13) does not hold and (10.15) holds. By symmetry this case is completely analogous to Case 2. □

10.4 Two examples with three stations

In this section, we consider the case $J=3$ and give two examples which show that λ_C, μ_i and p_{iC}, $i\in\mathcal{J}, C\in\mathcal{P}$, are not sufficient to determine the stability of the system. The first example shows that the stability of the system may depend on the tie-breaking rule $\gamma_{C,i}, C\in\mathcal{P}, i\in\mathcal{J}$. The second example shows that the stability of the system may depend not only on the mean service times but also on the distributions of the service times.

Both examples fit into a class of networks, whose structure is pictured in Figure 10.1.

The network has three stations, each represented by a circle. Each station serves customers in its queue, which is represented by an open rectangle. In each example, there are potentially four types of exogenous arrival processes, which are assumed to be four independent Poisson processes. The first three processes correspond to arrivals which are dedicated to queues 1, 2, and 3 respectively. The fourth Poisson process corresponds to arrivals which join the shorter of the two queues 1 and 2. If the queue lengths are equal at the time of an arrival, the customer breaks the tie using a Bernoulli(r) random variable which is independent of all the other primi-

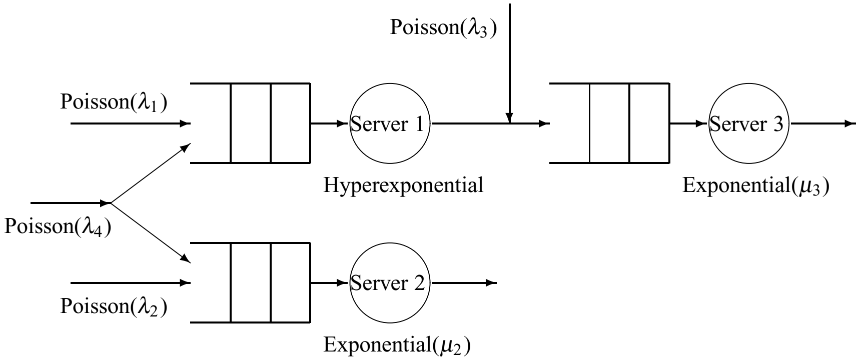


Fig. 10.1: A JSQ network

tive processes, with a success indicating that the customer joins queue 1. The four Poisson processes have rates λ_i, $i = 1,2,3,4$.

The service times at stations 2 and 3 are assumed to be i.i.d. exponential random variables with rates μ_2 and μ_3, respectively. The service times at station 1 are assumed to be i.i.d. random variables which are hyperexponential. We assume that the hyperexponential is generated by mixing independent *exponential*(a) and *exponential*(b) random variables, with the first component being chosen with probability ν. With these assumptions, the natural definition of the rate of service at station 1 is then:

$$\mu_1 = (\nu a^{-1} + (1-\nu)b^{-1})^{-1}.$$

Now, in such a network let

$$Y(t) = \begin{cases} 0, & \text{if no job is in service at station 1 ;} \\ 1, & \text{if the current job in service at station 1} \\ & \text{is assigned an } \mathit{exponential}(a) \text{ service;} \\ 2, & \text{if the current job in service at station 1} \\ & \text{is assigned an } \mathit{exponential}(b) \text{ service.} \end{cases}$$

Then, for this class of networks both $\{(Z_1(t), Z_2(t), Y(t)) : t \geq 0\}$ and $\{(Z_1(t), Z_2(t), Z_3(t), Y(t)), t \geq 0\}$ are irreducible continuous time Markov chains (CTMCs). When

$$\lambda_1 < \mu_1,\ \lambda_2 < \mu_2 \text{ and } \lambda_1 + \lambda_2 + \lambda_4 < \mu_1 + \mu_2, \tag{10.1}$$

it follows from Theorem 10.3.1 that the continuous time Markov chain $\{(Z_1(t), Z_2(t), Y(t)) : t \geq 0\}$ is positive recurrent. We use $\mathbb{P}\{(Z_1(\infty), Z_2(\infty)) \in \cdot\}$ to denote the stationary distribution of $\{(Z_1(t), Z_2(t)) : t \geq 0\}$. Recall that $A_1(t)$ is the number of customers that have entered either the queue or service at station 1 in $[0,t]$, and that $D_1(t)$ is the number of service completions by station 1 in $[0,t]$. Note that $A_1(t)/t$ and $D_1(t)/t$ are the arrival rate at station 1 the departure rate from station 1, respectively, in $[0,t]$. For a fixed time t, both of these rates are random. Our next proposition shows that, when (10.1) is satisfied, these rates converge to constants as $t \longrightarrow \infty$.

Property 10.4.1 *Assume that condition (10.1) holds.*

(a) Set $d_1 = \mu_1 \mathbb{P}\{Z_1(\infty) > 0\}$. *For each initial state x,*

$$\mathbb{P}_x \left\{ \lim_{t\longrightarrow\infty} D_1(t)/t = d_1 \right\} = 1. \tag{10.2}$$

(b) Set $a_1 = \lambda_1 + \lambda_4 \Big(\mathbb{P}\{Z_1(\infty) < Z_2(\infty)\} + r\mathbb{P}\{Z_1(\infty) = Z_2(\infty)\} \Big)$. *For each initial state x,*

$$\mathbb{P}_x \left\{ \lim_{t\longrightarrow\infty} A_1(t)/t = a_1 \right\} = 1. \tag{10.3}$$

(c) $a_1 = d_1$.

Proof. The proofs of both (a) and (b) follow by applying standard sample path versions of PASTA as in Wolff [15] Chapter 5, Theorem 6 and Example 5-23. We outline the proof for (a), the proof for (b) uses similar arguments. All arguments hold for the probability measure generated by a fixed, but arbitrary initial state x.

Let $\{N(t), t \geq 0\}$ be a Poisson process with rate μ_1. This process generates departures from station 1 whenever there is a job present at the station, otherwise an event in $N(\cdot)$ is ignored. Recall that $D_1(t)$ is the number of departures from station 1 in $[0,t]$. Then sample path PASTA and standard results for ergodic CTMC's yield:

$$\lim_{t\longrightarrow\infty} \frac{D_1(t)}{N(t)} = \mathbb{P}\{Z_1(\infty) > 0\} \qquad \text{a.s.}$$

The strong law of large numbers for renewal processes gives:

$$\lim_{t\longrightarrow\infty} \frac{N(t)}{t} = \mu_1 \qquad \text{a.s.}$$

Thus

$$\lim_{t\longrightarrow\infty} \frac{D_1(t)}{t} = \frac{D_1(t)}{N(t)} \frac{N(t)}{t} = \mu_1 \mathbb{P}\{Z_1(\infty) > 0\} \qquad \text{a.s.}$$

To prove (c), we note that from the proof of Theorem 10.3.1, the fluid model of the network consisting of the first two queues is stable. Thus, the network is rate stable, see for example, Dai [4]. Rate stability implies that $d_1 = a_1$, proving part (c).

When condition (10.1) holds, Proposition 10.4.1 asserts that the long-run departure rate from station 1 exists and is equal to d_1, a component of our next proposition. Note that $Z(t) = (Z_1(t), Z_2(t), Z_3(t))$ for the 3-station network of this section.

Property 10.4.2 *For the network in Figure 10.1, the Markov chain* $\{(Z(t),Y(t)) : t \geq 0\}$ *is positive recurrent iff*

$$\lambda_1 < \mu_1,\ \lambda_2 < \mu_2,\ \lambda_1+\lambda_2+\lambda_4 < \mu_1+\mu_2 \quad \textit{and} \quad \lambda_3+d_1 < \mu_3.$$

To prove Proposition 10.4.2, we first state and prove the following lemma, as applied to the network in Figure 10.1. Clearly, the lemma can be extended to a general setting like multiclass queueing networks with general distributions as in Dai [2] or stochastic processing networks as in Dai and Lin [5].

Lemma 10.4.3 *Assume that the continuous time Markov chain* $\{(Z(t),Y(t)) : t \geq 0\}$ *is positive recurrent with stationary distribution* $\pi = \{\pi_{i_1,i_2,i_3,i_4} : (i_1,i_2,i_3,i_4) \in \mathbb{Z}_+^4\}$. *Let the initial state* $(Z(0),Y(0)) = x$ *be fixed. Then,* $\mathbb{P}_x$*-a.s., for each fluid limit* $((\bar{T}_1,\bar{T}_2,\bar{T}_3),(\bar{Z}_1,\bar{Z}_2,\bar{Z}_3))$,

$$\bar{T}_j(t) = \left(1 - \sum_{(i_1,i_2,i_3,i_4)\in\mathbb{B}_j} \pi_{(i_1,i_2,i_3,i_4)}\right) t \tag{10.4}$$

for each $j = 1,2,3$ *and each* $t \geq 0$, *where* $\mathbb{B}_j = \{(i_1,i_2,i_3,i_4) \in \mathbb{Z}_+^4 : i_j = 0\}$.

Proof. For notational convenience, we prove the case for $j = 1$. The proofs for other cases are identical.

Since a nonidling service policy is assumed, we have for each $s \geq 0$

$$\frac{T_1(s)}{s} = \frac{1}{s}\int_0^s \mathbf{1}_{\{Z_1(u)>0\}}\,du = 1 - \frac{1}{s}\int_0^s \mathbf{1}_{\{Z_1(u)=0\}}\,du.$$

By the positive recurrence of the Markov chain, we have

$$\begin{aligned}\mathbb{P}_x\Bigg\{\lim_{s\to\infty}\frac{T_1(s)}{s} &= 1 - \lim_{s\to\infty}\frac{1}{s}\int_0^s \mathbf{1}_{\{Z_1(u)=0\}}\,du \\ &= 1 - \sum_{(i_1,i_2,i_3,i_4)\in\mathbb{B}_1} \pi_{(i_1,i_2,i_3,i_4)}\Bigg\} = 1.\end{aligned} \tag{10.5}$$

For each sample path in the event set of (10.5) and for each $t \geq 0$,

$$\begin{aligned}\bar{T}_1(t) &= \lim_{n\to\infty}\frac{T_1(nt)}{n} \\ &= t\lim_{s\to\infty}\frac{T_1(s)}{s} \\ &= t\left(1 - \sum_{(i_1,i_2,i_3,i_4)\in\mathbb{B}_1} \pi_{(i_1,i_2,i_3,i_4)}\right),\end{aligned}$$

thus proving the lemma. □

Proof of Proposition 10.4.2 Recall that the 3-dimensional process $\{(Z_1(t), Z_2(t), Y(t)) : t \geq 0\}$ is an irreducible CTMC. If $\lambda_1 \geq \mu_1$ or $\lambda_2 \geq \mu_2$ or $\lambda_1 + \lambda_2 + \lambda_4 \geq \mu_1 + \mu_2$, then by Theorem 10.3.1, the 3-dimensional CTMC is not positive recurrent, and so neither is the 4-dimensional CTMC $\{(Z_1(t), Z_2(t), Z_3(t), Y(t)) : t \geq 0\}$. This establishes the necessity of the first three conditions in Proposition 10.4.2.

Thus, we assume that $\lambda_1 < \mu_1$, $\lambda_2 < \mu_2$ and $\lambda_1 + \lambda_2 + \lambda_4 < \mu_1 + \mu_2$ throughout the remainder of this proof. Let $\{\kappa_{ijk}(r) : i, j, k\}$ be the stationary distribution of the 3-dimensional Markov chain $\{(Z_1(t), Z_2(t), Y(t)) : t \geq 0\}$. We now show that $\{(Z_1(t), Z_2(t), Z_3(t), Y(t)) : t \geq 0\}$ is positive recurrent if and only if

$$\lambda_3 + \mu_1 \left(1 - \sum_{j=0}^{\infty} \sum_{k=0}^{2} \kappa_{0jk}(r) \right) < \mu_3. \tag{10.6}$$

Fix an initial state $(Z(0), Y(0))$, say, $(Z(0), Y(0)) = (0,0,0,0)$. Let $((\bar{T}_1, \bar{T}_2, \bar{T}_3), (\bar{Z}_1, \bar{Z}_2, \bar{Z}_3))$ be a fluid limit. It follows that it satisfies the following fluid model equation (see, e.g., Dai [2])

$$\bar{Z}_3(t) = \lambda_3 t + \mu_1 \bar{T}_1(t) - \mu_3 \bar{T}_3(t). \quad t \geq 0,$$

Applying Lemma 10.4.3 to the 3-dimensional Markov chain, we have

$$\bar{Z}_3(t) = \left[\lambda_3 + \mu_1 \Big(1 - \sum_{j,k} \kappa_{0jk}(r) \Big) \right] t - \mu_3 \bar{T}_3(t), \quad t \geq 0. \tag{10.7}$$

Assume that $\{(Z_1(t), Z_2(t), Z_3(t), Y(t)) : t \geq 0\}$ is positive recurrent with stationary distribution $\pi = \{\pi_{(i_1,i_2,i_3,i_4)}\}$, but that condition (10.6) does not hold. Since $\sum_{(i_1,i_2,i_4) \in \mathbb{Z}_+^3} \pi_{(i_1,i_2,0,i_4)} > 0$, it follows from Lemma 10.4.3 and (10.7) that $\bar{Z}_3(t) > 0$ for each fluid limit and each time $t > 0$. Therefore, the fluid limit model is weakly unstable as defined in [3]. It follows from Theorem 4.2 of [3] that $\{(Z_1(t), Z_2(t), Z_3(t), Y(t)) : t \geq 0\}$ is transient, and hence not positive recurrent, contradicting the assumption that $\{(Z_1(t), Z_2(t), Z_3(t), Y(t)) : t \geq 0\}$ is positive recurrent. Thus we have proved that $\{(Z_1(t), Z_2(t), Z_3(t), Y(t)) : t \geq 0\}$ is positive recurrent only if (10.6) holds.

Now suppose that (10.6) holds. For each fluid limit $((\bar{T}_1, \bar{T}_2, \bar{T}_3), (\bar{Z}_1, \bar{Z}_2, \bar{Z}_3))$, we have $\bar{Z}_3(t) \geq 0$ for each $t \geq 0$. Thus, (10.7) implies that

$$\bar{T}_3(t) \leq \frac{\lambda_3 + \mu_1 (1 - \sum_{j,k} \kappa_{0jk}(r))}{\mu_3} \cdot t = (1 - \varepsilon) t, \tag{10.8}$$

where

$$\varepsilon = 1 - \frac{1}{\mu_3} \left(\lambda_3 + \mu_1 \left(1 - \sum_{j,k} \kappa_{0jk}(r) \right) \right) \ > \ 0.$$

Since (10.8) holds for every fluid limit, we have, $\mathbb{P}_x$-a.s.,

$$\limsup_{t\to\infty}\frac{1}{t}T_3(t) \le 1-\varepsilon.$$

Therefore, $\mathbb{P}_x$-a.s.,

$$\liminf_{t\to\infty}\frac{1}{t}\int_0^t \mathbf{1}_{\{Z_3(u)=0\}}du \ge \varepsilon. \tag{10.9}$$

Let B be a finite set such that

$$\sum_{(i,j,k)\notin B} \kappa_{ijk}(r) < \varepsilon. \tag{10.10}$$

Now, define $\tilde{B} \equiv \{(i,j,0,k) : (i,j,k) \in B\}$. By (10.9) and (10.10), $\mathbb{P}_x$-a.s.,

$$\begin{aligned}
&\liminf_{t\to\infty}\frac{1}{t}\int_0^t \mathbf{1}_{\{(Z_1(u),Z_2(u),Z_3(u),Y(u))\in\tilde{B}\}}du \\
&\ge \liminf_{t\to\infty}\frac{1}{t}\int_0^t \mathbf{1}_{\{Z_3(u)=0\}}du - \lim_{t\to\infty}\frac{1}{t}\int_0^t \mathbf{1}_{\{(Z_1(u),Z_2(u),Y(u))\notin B\}}du \\
&= \liminf_{t\to\infty}\frac{1}{t}\int_0^t \mathbf{1}_{\{Z_3(u)=0\}}du - \sum_{(i,j,k)\notin B} \kappa_{ijk}(r) \\
&\ge \varepsilon - \sum_{(i,j,k)\notin B} \kappa_{ijk}(r) > 0.
\end{aligned}$$

By Fatou's lemma and Fubini's theorem,

$$\begin{aligned}
&\liminf_{t\to\infty}\frac{1}{t}\int_0^t \mathbb{P}_x\Big\{(Z_1(u),Z_2(u),Z_3(u),Y(u)) \in \tilde{B}\Big\}du \\
&\ge \varepsilon - \sum_{(i,j,k)\notin B} \kappa_{ijk}(r) > 0.
\end{aligned} \tag{10.11}$$

Since $\tilde{B}$ is a finite set, (10.11) implies that the Markov chain $\{(Z_1(t),Z_2(t),Z_3(t), Y(t)) : t \ge 0\}$ is positive recurrent. □

We are now ready to analyze a set of examples which give further insight into the stability behavior of JSQ networks.

Example 1. We now consider a special case of three station network introduced above. Let $\lambda_1 = \lambda_2 = 0$ and let λ_3 and λ_4 be arbitrary. Furthermore, assume $\nu = \frac{1}{2}$ and $\mu_1 := a = b = \mu_2$. Thus, there are exponential service times at all stations, with stations 1 and 2 having the same service rates. The 3-dimensional process $\{(Z_1(t),Z_2(t),Z_3(t)) : t \ge 0\}$ is a Markov process. Further it is positive recurrent if and only the 4-dimensional Markov process $\{(Z_1(t),Z_2(t),Z_3(t),Y(t)) : t \ge 0\}$ is positive recurrent.

We now argue that the positive recurrence of $\{(Z_1(t),Z_2(t),Z_3(t)) : t \ge 0\}$ depends on the tie-breaking parameter r. The first three conditions in Proposition 10.4.2 reduces to $\lambda_4 < 2\mu_1$. Under this condition, by Theorem 10.3.1, the process

$\{(Z_1(t), Z_2(t)) : t \geq 0\}$ is positive recurrent. Let $\{\kappa_{ij}(r) : i, j \geq 0\}$ be the stationary distribution of this process. Then applying Proposition 10.4.2 we immediately obtain:

Claim. If $\lambda_4 \geq 2\mu_1$ then the Markov chain $\{(Z_1(t), Z_2(t), Z_3(t)) : t \geq 0\}$ is not positive recurrent. If $\lambda_4 < 2\mu_1$, then $\{(Z_1(t), Z_2(t), Z_3(t)) : t \geq 0\}$ is positive recurrent if and only if

$$\lambda_3 + \mu_1 \left(1 - \sum_{j=0}^{\infty} \kappa_{0j}(r) \right) < \mu_3. \tag{10.12}$$

By Lemma 7 in Dia, Hasenbein and Kim [6], it is seen that $\sum_{j=0}^{\infty} \kappa_{0j}(r)$ decreases strictly as r increases. Thus it is clear that one can choose fixed parameters $\lambda_3, \lambda_4, \mu_1$, and μ_3 for which the stability conditions will hold for some r and not hold for another choice of r. *In particular, the necessary and sufficient conditions for the positive recurrence of X depend on the tie-breaking parameter r.*

Example 2. Consider now another special case of the network depicted in Figure 10.1. In particular let $\lambda_1 = \lambda_2 = 0.8$, $\lambda_3 = 0.17$, $\lambda_4 = 0.1$, $r = 1/2$ and $\mu_1 = \mu_2 = \mu_3 = 1$, where station 1's service time distribution remains to be chosen. We now argue that the positive recurrence of $\{(Z_1(t), Z_2(t), Z_3(t), Y(t)) : t \geq 0\}$ depends on the distribution of the service times for station 1 even if the mean is fixed.

First suppose $\nu = \frac{1}{2}$ and $a = b = 1$. Thus, all service times for station 1 are exponentially distributed with mean 1. For this case, we have the following claim:

Claim. If the service times for station 1 are exponentially distributed with mean 1 then the process $\{(Z_1(t), Z_2(t), Z_3(t)) : t \geq 0\}$ is not positive recurrent.

Proof of Claim 10.4 For the set of parameters under consideration, condition (10.1) holds and we can apply Proposition 10.4.1, which implies that the departure rate from station 1 (and station 2) exists with probability 1. As argued earlier, from Theorem 10.3.1, condition (10.1) also implies that the fluid model of the network consisting of the first two queues is stable, and so the network itself is rate stable. Hence, so the total departure rate from the first two queues must equal the total arrival rate of 1.7. Furthermore, by symmetry, the departures rates from station 1 and station 2 must be equal. Thus, $d_1 = 0.85$ and applying Proposition 10.4.2, we infer that $\{(Z_1(t), Z_2(t), Z_3(t)) : t \geq 0\}$ is not positive recurrent. □

Now suppose we alter the distribution, but not the mean service time, at station 1. In particular, let $0 < \nu < 1$ and $a = \frac{\nu}{1-\nu+\nu^2}$ and $b = 1/\nu$. Then services at station 1 are hyperexponential with the following c.d.f.: For $0 \leq x < \infty$,

$$F(x) = \nu \left(1 - \exp\left(\frac{-\nu x}{1 - \nu + \nu^2} \right) \right) + (1 - \nu) \left(1 - \exp(-\frac{x}{\nu}) \right). \tag{10.13}$$

Note that for any $0 < \nu < 1$ the mean service time is 1.

Claim. If the service times for station 1 are hyperexponential as described above and if v is sufficiently small, then the process $\{(Z_1(t), Z_2(t), Z_3(t), Y(t)) : t \geq 0\}$ is positive recurrent.

Proof of Claim 10.4 In this case, Proposition 10.4.1 gives:

$$d_1 = 0.8 + 0.1 \left[\mathbb{P}\{Z_1(\infty) < Z_2(\infty)\} + 0.5\mathbb{P}\{Z_1(\infty) = Z_2(\infty)\}\right]. \tag{10.14}$$

Then, by Proposition 10.4.2, X is positive recurrent iff

$$0.17 + 0.8 + 0.1\mathbb{P}\{Z_1(\infty) < Z_2(\infty)\} + 0.05\mathbb{P}\{Z_1(\infty) = Z_2(\infty)\} < 1,$$

or equivalently

$$10\mathbb{P}\{Z_1(\infty) < Z_2(\infty)\} + 5\mathbb{P}\{Z_1(\infty) = Z_2(\infty)\} < 3.$$

A sufficient condition for the inequality above to hold is

$$g(v) \equiv \mathbb{P}\{Z_1(\infty) \leq Z_2(\infty)\} \ < \ 0.3. \tag{10.15}$$

We will show that this is true for v sufficiently small. Observe that

$$Z_1(\infty) \ \geq_{\text{st}} \ Z^{v}_{M/G/1} \ \text{ and } \ Z_2(\infty) \ \leq_{\text{st}} \ Z_{M/M/1}, \tag{10.16}$$

where $Z^{v}_{M/G/1}$ denotes a random variable whose distribution is the stationary distribution of the number of customers in an ordinary $M/G/1$ queue with arrival rate 0.8 and service time distribution function F given by (10.13), and $Z_{M/M/1}$ denotes a random variable whose distribution is the stationary distribution of the number of customers in an ordinary $M/M/1$ queue with arrival rate 0.9 and service rate 1.

Since the Laplace-Stieltjes transform (LST) of service times in the $M/G/1$ queue is

$$\int_0^{\infty} e^{-sx} dF(x) = \frac{v^2}{(1-v+v^2)s+v} + \frac{1-v}{sv+1}, \qquad Re(s) > 0,$$

the Pollaczek-Khintchine (see, e.g., p. 260 in [12]) formula yields

$$\mathrm{e}[z^{Z^{v}_{M/G/1}}] = \frac{(4-3v+8v^2) - 4(1-2v+2v^2)z}{5(4-3v+8v^2) - 4(5-v+6v^2+4v^3)z + 16v(1-v+v^2)z^2},$$

which is the probability generating function for the number of customers in the $M/G/1$ queue, in stationarity. Therefore

$$\lim_{v \to 0+} \mathrm{e}[z^{Z^{v}_{M/G/1}}] = 0.2 \quad |z| < 1.$$

Now applying the continuity theorem for probability generating functions (c.f. Theorem 1.5.1 in [13]) we have

$$\lim_{v \to 0+} \mathbb{P}(Z^{v}_{M/G/1} \leq x) = 0.2 \quad \text{for all} \quad 0 < x < \infty. \tag{10.17}$$

By (10.16),

$$\begin{aligned} g(\nu) &= 1 - \mathbb{P}\{Z_1(\infty) > Z_2(\infty)\} \\ &\le 1 - \mathbb{P}\{Z_1(\infty) > x > Z_2(\infty)\} \\ &= 1 - \mathbb{P}(\{Z_2(\infty) < x\} - \{Z_1(\infty) \le x\}) \\ &\le 1 - \mathbb{P}\{Z_2(\infty) < x\} + \mathbb{P}\{Z_1(\infty) \le x\} \\ &\le 1 - \mathbb{P}(Z_{M/M/1} < x) + \mathbb{P}(Z^{\nu}_{M/G/1} \le x), \end{aligned}$$

for any $0 < x < \infty$. Hence, by (10.17),

$$\limsup_{\nu \to 0+} g(\nu) \le 1.2 - \mathbb{P}(Z_{M/M/1} < x).$$

Letting $x \longrightarrow \infty$ leads to

$$\limsup_{\nu \to 0+} g(\nu) \le 0.2.$$

Therefore (10.15) holds for sufficiently small ν. Hence, for sufficiently small ν, $\{(Z_1(t), Z_2(t), Z_3(t), Y(t)) : t \ge 0\}$ is positive recurrent when the service times for station 1 have the hyperexponential distribution function (10.13). □

Claims 2 and 3, taken together, show that the positive Harris recurrence of X depends on more than just the mean values of the primitive distributions in the network.

10.5 JSQ networks with homogeneous feedback

As we see in Section 10.3 the stability of a 2-station network does not depend on the distributions of interarrival and service times or the tie-breaking probabilities. Unfortunately, when $J \ge 3$, the analogous result does not hold as we see in Section 10.4. However, for such networks we can identify stability conditions in terms of λ_C, μ_C and $p_{iC}, i \in \mathcal{J}, C \in \mathcal{P}$, under an additional assumption on network structure.

Assumption 10.5.1 *For any $C \in \mathcal{P}$, p_{iC} does not depend on $i \in \mathcal{J}$. Namely, all stations have the same feedback probabilities. For $C \in \mathcal{P}$, let*

$$\Lambda_C \equiv \sum_{B:\phi \ne B \subseteq C} \lambda_B, \; P_C \equiv \sum_{B:\phi \ne B \subseteq C} p_{iB} \text{ and } \mu_C \equiv \sum_{i \in C} \mu_i.$$

Let $\lambda^* \equiv \Lambda_{\mathcal{J}}$ be the total external arrival rate to the network and $p^* \equiv p^*_i$, which is independent of station $i \in \mathcal{J}$. To avoid triviality, further assume that $p^* < 1$.

Under this assumption, the stability of larger networks can be determined directly from the first-order network parameters, as the following result demonstrates.

Theorem 10.5.2 *Consider a JSQ network with $J \geq 3$ whose parameters are in concordance with Assumption 1. The Markov process $\{X(t) : t \geq 0\}$ is positive Harris recurrent if and only if*

$$\Lambda_C + \frac{\lambda^*}{1-p^*} P_C < \mu_C, \ \text{for all } C \in \mathcal{P}. \tag{10.1}$$

To prove the theorem, we first need to prove the following lemma.

Lemma 10.5.3 *Let $\bar{\mathbb{X}}$ be a fluid model solution. Consider a fixed regular $t > 0$ and let $C \equiv C(t) = \{i \in \mathcal{J} : \bar{Z}_i(t) > 0\}$. Then*

$$\sum_{i \in C} \dot{\bar{Z}}_i(t) = \frac{1-p^*}{1-p^*+P_C} \left(\Lambda_C + \frac{\lambda^*}{1-p^*} P_C - \mu_C \right). \tag{10.2}$$

Proof. Using (10.21), we have

$$\sum_{i \in \mathcal{J}} \dot{\bar{A}}_i(t) = \lambda^* + p^* \sum_{i \in \mathcal{J}} \mu_i \dot{\bar{T}}_i(t) \tag{10.3}$$

and

$$\sum_{i \in C} \dot{\bar{A}}_i(t) = \Lambda_C + P_C \sum_{i \in \mathcal{J}} \mu_i \dot{\bar{T}}_i(t). \tag{10.4}$$

Subtracting (10.4) from (10.3), yields

$$\sum_{i \in \mathcal{J}-C} \dot{\bar{A}}_i(t) = \lambda^* - \Lambda_C + (p^* - P_C) \sum_{i \in \mathcal{J}} \mu_i \dot{\bar{T}}_i(t). \tag{10.5}$$

Since $\bar{Z}_i(t) = 0$ for $i \in \mathcal{J} - C$, $\dot{\bar{Z}}_i(t) = 0$ for $i \in \mathcal{J} - C$. Hence, by (10.15),

$$\dot{\bar{A}}_i(t) = \mu_i \dot{\bar{T}}_i(t), \ \ i \in \mathcal{J} - C. \tag{10.6}$$

Then (10.5) and (10.6) give

$$\sum_{i \in \mathcal{J}-C} \mu_i \dot{\bar{T}}_i(t) = \frac{\lambda^* - \Lambda_C + (p^* - P_C) \sum_{i \in C} \mu_i \dot{\bar{T}}_i(t)}{1-p^*+P_C}. \tag{10.7}$$

Since $\bar{Z}_i(t) > 0$, (10.18) and (10.19) imply

$$\dot{\bar{T}}_i(t) = 1, \ \ i \in C. \tag{10.8}$$

Substituting (10.7) and (10.8) into (10.4) leads to

$$\sum_{i \in C} \dot{\bar{A}}_i(t) = \frac{(1-p^*)\Lambda_C + \lambda^* P_C + \mu_C P_C}{1-p^*+P_C}. \tag{10.9}$$

Next, (10.15) and (10.8) give us

$$\sum_{i\in C} \dot{\bar{Z}}_i(t) = \sum_{i\in C} \dot{\bar{A}}_i(t) - \mu_C. \tag{10.10}$$

Finally, substituting (10.9) into (10.10) yields (10.2).

Proof of Sufficiency of Theorem 10.5.2 Suppose $\bar{\mathbb{X}}$ is a fluid model solution. Let $f(t) = |\bar{Z}(t)|$. Consider a fixed regular $t > 0$ with $f(t) > 0$ and again let $C = \{i \in \mathcal{J} : \bar{Z}_i(t) > 0\}, t \geq 0$. Since $\dot{\bar{Z}}_i(t) = 0$ for $i \in \mathcal{J} - C$, $\dot{f}(t) = \sum_{i\in C} \dot{\bar{Z}}_i(t)$. Hence, by Lemma 10.5.3,

$$\dot{f}(t) \leq -\varepsilon, \tag{10.11}$$

for any such t, where

$$\varepsilon = \min_{B\in\mathcal{P}} \frac{1-p^*}{1-p^*+P_B}\left(\mu_B - \Lambda_B - \frac{\lambda^*}{1-p^*}P_B\right) > 0.$$

From (10.11), it is readily seen that the fluid model is stable. The proof is now completed by applying Lemma 10.2.2. □

Proof of Necessity of Theorem 10.5.2: By Lemma 10.2.5, it suffices to show that the augmented fluid model is weakly unstable if (10.1) does not hold for some $C \in \mathcal{P}$. Suppose then that (10.1) does not hold for some $C \in \mathcal{P}$. Let $\bar{\mathbb{X}}$ be an augmented fluid model solution with $\bar{Z}(0) = 0$. In light of (10.15) and (10.20) we have

$$\sum_{i\in C} \dot{\bar{Z}}_i(t) \geq \Lambda_C + P_C \sum_{i\in\mathcal{J}} \mu_i \dot{\bar{T}}_i(t) - \sum_{i\in C} \mu_i \dot{\bar{T}}_i(t). \tag{10.12}$$

Next, using (10.15) and (10.21), $\sum_{i\in\mathcal{J}} \dot{\bar{Z}}_i(t) = \lambda^* + (p^*-1)\sum_{i\in\mathcal{J}} \mu_i \dot{\bar{T}}_i(t)$, which can be rewritten as

$$\sum_{i\in\mathcal{J}} \mu_i \dot{\bar{T}}_i(t) = \frac{\lambda^*}{1-p^*} - \frac{1}{1-p^*}\sum_{i\in\mathcal{J}} \dot{\bar{Z}}_i(t). \tag{10.13}$$

Substituting (10.13) into (10.12) yields

$$\sum_{i\in C} \dot{\bar{Z}}_i(t) \geq \Lambda_C + \frac{\lambda^*}{1-p^*}P_C - \frac{P_C}{1-p^*}\sum_{i\in\mathcal{J}} \dot{\bar{Z}}_i(t) - \sum_{i\in C} \mu_i \dot{\bar{T}}_i(t).$$

Equation (10.23) then implies that

$$\frac{1-p^*+P_C}{1-p^*}\sum_{i\in C} \dot{\bar{Z}}_i(t) + \frac{P_C}{1-p^*}\sum_{i\in\mathcal{J}-C} \dot{\bar{Z}}_i(t) > \Lambda_C + \frac{\lambda^*}{1-p^*}P_C - \mu_C. \tag{10.14}$$

Now, let

$$f(t)=\frac{1-p^*+P_C}{1-p^*}\sum_{i\in C}\bar{Z}_i(t)+\frac{P_C}{1-p^*}\sum_{i\in\mathcal{J}-C}\bar{Z}_i(t).$$

Then by (10.14) and the negation of (10.1), $\dot{f}(t)>0$ for all $t>0$, which proves that the augmented fluid model is weakly unstable. □

The following theorem provides a sufficient condition for the Markov process X to be unstable in the sense that $|Z(t)|\longrightarrow\infty$ as $t\longrightarrow\infty$ with probability 1.

Theorem 10.5.4 *Consider a JSQ network with $J\geq 3$ whose parameters are in concordance with Assumption 1. The process X is unstable in the sense that $|Z(t)|\longrightarrow\infty$ as $t\longrightarrow\infty$ with probability 1 if there exists a $C\in\mathcal{P}$ such that*

$$\Lambda_C+\frac{\lambda^*}{1-p^*}P_C>\mu_C. \tag{10.15}$$

To prove the theorem, we first need the following lemma.

Lemma 10.5.5 *Let $\bar{\mathbb{X}}$ be a fluid model solution. Then, for any $C\in\mathcal{P}$,*

$$\frac{1-p^*+P_C}{1-p^*}\sum_{i\in C}\dot{\bar{Z}}_i(t)+\frac{P_C}{1-p^*}\sum_{i\in\mathcal{J}-C}\dot{\bar{Z}}_i(t)\geq\Lambda_C+\frac{\lambda^*}{1-p^*}P_C-\mu_C. \tag{10.16}$$

Proof. Equations (10.15) and (10.20) imply,

$$\sum_{i\in C}\dot{\bar{Z}}_i(t)\geq\Lambda_C+P_C\sum_{i\in\mathcal{J}}\mu_i\dot{\bar{T}}_i(t)-\sum_{i\in C}\mu_i\dot{\bar{T}}_i(t). \tag{10.17}$$

Now (10.15) and (10.21) give $\sum_{i\in\mathcal{J}}\dot{\bar{Z}}_i(t)=\lambda^*+(p^*-1)\sum_{i\in\mathcal{J}}\mu_i\dot{\bar{T}}_i(t)$, which can be rewritten as

$$\sum_{i\in\mathcal{J}}\mu_i\dot{\bar{T}}_i(t)=\frac{\lambda^*}{1-p^*}-\frac{1}{1-p^*}\sum_{i\in\mathcal{J}}\dot{\bar{Z}}_i(t). \tag{10.18}$$

By substituting (10.18) into (10.17), we get

$$\sum_{i\in C}\dot{\bar{Z}}_i(t)\geq\Lambda_C+\frac{\lambda^*}{1-p^*}P_C-\frac{P_C}{1-p^*}\sum_{i\in\mathcal{J}}\dot{\bar{Z}}_i(t)-\sum_{i\in C}\mu_i\dot{\bar{T}}_i(t).$$

Since $\dot{\bar{T}}_i(t)\leq 1$, $i\in C$, (10.16) is obtained. □

Proof of Theorem 10.5.4: Suppose $\bar{\mathbb{X}}$ is a fluid model solution with $\bar{Z}(0)=0$, and let $C\in\mathcal{P}$ be such that it satisfies (10.15). Let

$$f(t)=\frac{1-p^*+P_C}{1-p^*}\sum_{i\in C}\bar{Z}_i(t)+\frac{P_C}{1-p^*}\sum_{i\in\mathcal{J}-C}\bar{Z}_i(t).$$

By Lemma 10.5.5, $\dot{f}(t) > 0$ for all $t > 0$. Thus $f(t) > 0$ and so $|\bar{Z}(t)| > 0$ for all $t > 0$. Hence the fluid model is weakly unstable and the proof is completed by applying Lemma 10.2.3. □

10.6 Further study

In this section, we briefly mention some further research topics. A direct extension of the JSQ network studied in this chapter is a multiserver JSQ network. In such a network each station has one or more identical servers, each with a dedicated buffer for waiting customers. When customers enter the system, or complete processing at a station, they choose some subset of stations, and buffers, in the network and join the shortest queue. The techniques used in this chapter can be readily extended to multiserver JSQ networks. Stability results for such networks are of particular interest due to intriguing conjectures put forth by Suhov and Vvedenskaya [14]. In fact, these conjectures can be extended, and resolved (positively) using the techniques in this chapter.

Another extension to the framework of this chapter is a JSQ network with state dependent service rates. In this network the service rate of each server may depend on the states (e.g., busy or idle) of the other servers in the network. Such a network has been of interest due to applications in wireless networks and there has been some progress in obtaining stability conditions for single station networks of this type. It is likely that the techniques in this chapter can also be extended to JSQ networks with state dependent service rates.

Acknowledgements J.G. Dai's research was supported in part by NSF grants CMMI-0727400 and CMMI-0825840. John Hasenbein was supported by NSF grant CMMI-0800676 during the course of this research. Bare Kim's research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency), (NIPA-2009-2009-C1090-0902-0013).

References

1. Bramson, M.: Stability of two families of queueing networks and a discussion of fluid limits. Queueing Syst. Theory and Appl. **28**, 7–31 (1998)
2. Dai, J.G.: On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. Ann. Appl. Probab. **5**, 49–77 (1995)
3. Dai, J.G.: A fluid-limit model criterion for instability of multiclass queueing networks. Ann. Appl. Probab. **6**, 751–757 (1996)
4. Dai, J.G.: Stability of fluid and stochastic processing networks. In: MaPhySto Miscellanea Publication, No. 9, Centre for Mathematical Physics and Stochastics (1999)
5. Dai, J.G., Lin, W.: Maximum pressure policies in stochastic processing networks. Oper. Res. **53**, 197–218 (2005)

6. Dai, J.G., Hasenbein, J.J., Kim, B.: Stability of join-the-shortest-queue networks. Queueing Syst. Theory and Appl. **57**, 129–145 (2007)
7. Foley, R.D., McDonald, D.R.: Join the Shortest Queue: Stability and Exact Asymptotics. Ann. Appl. Probab. **11**, 569–607 (2001)
8. Foss, S.G., Chernova, N.: On the stability of a partially accessible multi-station queue with state-dependent routing. Queueing Syst. Theory and Appl. **29** 55–73 (1998)
9. Jackson, J.R.: Networks of waiting lines. Oper. Res. **5**, 518–521 (1957)
10. Jackson, J.R.: Jobshop-like queueing systems. Manag. Sci. **10** 131–142 (1963)
11. Kurkova, I.A.: A load-balanced network with two servers. Queueing Syst. Theory and Appl. **37** 379–389 (2001)
12. Medhi, J.: Stochastic Models in Queueing Theory, 2nd edn. Academic, San Diego (2002)
13. Resnick, S.I.: Adventures in Stochastic Processes. Birkhäuser, Boston (1992)
14. Suhov, Y.M., Vvedenskaya, N.D.: Fast Jackson networks with dynamic routing. Probl. Inf. Transm. **38** 136–153 (2002)
15. Wolff, R.W.: Stochastic modeling and the theory of queues. Prentice Hall, Englewood Cliffs (1989)

Chapter 11
Methods in Diffusion Approximation for Multi-Server Systems: Sandwich, Uniform Attraction and State-Space Collapse

Hong Chen and Heng-Qing Ye

Abstract In this chapter, we demonstrate through simple queueing models some of the methods that have been developed for the diffusion approximation. Specifically, we first show how the sandwich method is used to establish the diffusion approximation for a multi-server queue, and next show how the uniform attraction and the state-space collapse method is used to establish the diffusion approximation for a multi-class queue under a first-in-first-out (FIFO) service discipline. Finally, we use all of the above methods to establish the diffusion approximation for a system with multi-channel queues to which the jobs are routed based on a join-the-shortest-queue (JSQ) routing control.

11.1 Introduction

There has been substantial literature on the diffusion approximation for a queueing system ever since the pioneer work of Kingman [13] and Iglehart and Whitt [10, 11]. The applications can be found to the modeling and analysis of manufacturing system, service system and and telecommunication networks. This chapter can be considered as a supplement to the book by Chen and Yao [4], from which the reader can find more references in the fluid and the diffusion approximations. The fluid and the diffusion approximations considered in Chen and Yao [4] are limited to the *single server* networks including the generalized Jackson networks, the feedforward multiclass networks and some general multiclass networks. In this chapter, we consider the fluid and the diffusion approximations for multi-server queues. In

Hong Chen
Sauder School of Business, University of British Columbia, Vancouver, Canada
e-mail: hong.chen@sauder.ubc.ca

Heng-Qing Ye
Faculty of Business, Hong Kong Polytechnic University
e-mail: lgtyehq@inet.polyu.edu.hk

R.J. Boucherie and N.M. van Dijk (eds.), *Queueing Networks: A Fundamental Approach*,
International Series in Operations Research & Management Science 154,
DOI 10.1007/978-1-4419-6472-4_11, © Springer Science+Business Media, LLC 2011

addition, we use different methodologies such as the sandwich method and the uniform attraction in establishing these approximations. In considering the fluid and the diffusion approximations, we assume that the number of servers remains constant in taking the limit. In contrast, there has been substantial research motivated from the study of call centers that considers the number of the servers growing to infinity in the limiting process. Readers are referred to the survey paper by Gans, et al. [9] and the references in Itay and Whitt [12] for this literature.

To the stochastic processes under consideration, the fluid approximation and the diffusion approximation resemble the strong law-of-large-numbers (SLLNs) and the central limit theorem (CLT) to the random sequences. Consider, for example, the summation, $X(n)$, of n independent and identically distributed random variables. The strong law-of-large-numbers suggests that $X(n)/n$ converge almost surely to a constant m (which is the common mean of the random variables), and the central limit theorem suggests that $\sqrt{n}[X(n)-m]$ converge weakly (or in distribution) to a normal distribution. Among others, these limiting results are fundamental to many applications; for example, SLLNs is fundamental to the point estimate and CLT is fundamental to the confidence interval in statistics. In studying the queueing systems, we are concerned with the dynamic evolution of the related processes (such as the queue length process and the workload process). With the above summation example, the fluid approximation is about the convergence of the fluid-scaled processes, $\bar{X}^n(t) := X(\lfloor nt \rfloor)/n$, as $n \to \infty$; when exists, its limit, denoted as $\bar{X}(t)$, is referred to as the fluid limit. The diffusion approximation (also referred to as the functional central limit theorem) is about the convergence of the diffusion-scaled processes, $\hat{X}^n(t) := \sqrt{n}[\bar{X}^n(t) - \bar{X}(t)]$ (sometimes n^2 is used in place of n on the right-hand-side), as $n \to \infty$; when exists, its limit, denoted as $\hat{X}(t)$, is referred to as the diffusion limit.[1] The Fluid limit is usually a deterministic process (included as the special cases are a linear process or a piecewise linear process) and the diffusion limit is usually a diffusion process (included as the special cases are the Brownian motion and the reflected Brownian motion). Both of these limits are much easier to characterize and to analyze than the original processes. Hence, they play the same role in studying the stochastic processes as the role that SLLNs and CLT play in statistics.

The standard procedure in establishing the diffusion approximation for a single class single server queueing system is through the use of the reflection mapping. The reflection mapping is used to uniquely characterize the dynamics of the queue length or the workload process and the cumulative idle time process. One of the key conditions in the reflection mapping is the dynamic complementarity condition. In the single class single server queue, this condition amounts to the non-idling service discipline; namely, the server never gets idle when there is at least one job in the system. This condition would fail; for example, in a multi-server queue, at least one of these servers will be idle when the number of jobs in the system is less than the number of the servers. However, when the number of jobs in the system is

[1] The convergence in the fluid approximation is usually the almost sure convergence, and the convergence in the diffusion approximation is usually the weak convergence for the stochastic process. More precise definitions of these modes of the convergence will be given in Section 2.

larger than the number of the servers, none of the servers will be idle. In this sense, the dynamic complementarity condition holds approximately. In particular, with the space scaling, the diffusion limit is expected to satisfy the dynamic complementarity condition and hence to be able to described by the reflected Brownian motion.

The reflection mapping may also fail to describe the dynamics of a multi-class queue or a system where one stream of jobs is routed to multiple queues upon arrivals. In this case, neither the queue length process nor the workload process can be uniquely described by the reflection mapping. However, under the appropriate condition, the total workload in the diffusion scale can be shown to converge to a reflected Brownian motion (which is described by the reflection mapping). The queue length process for each job class (in a multi-class queue) or for each queue (in a multiple queue system) in the diffusion limit is a constant multiple of the reflected Brownian motion. This phenomenon is known as the state-space collapse, in the sense that the queue length process, which is originally multi-dimensional, is reduced to a one-dimensional process in the diffusion limit (all proportional to a single reflected Brownian motion). A general framework for establishing the state-space collapse property is first to establish a uniform attraction property of the fluid limit. The uniform attraction means that given any fluid limit with a bounded initial state the fluid state will converge to a fixed point state as the time approaches infinity (very often in finite time). The key step to transform the uniform attraction property to the state-space collapse property involves a rescaling technique. Suppose now in the diffusion scaling of the concerned processes (such as the queue length process or the workload process), time is scaled by n^2 and space by n. By this technique, the order $O(n^2)$-long time interval of the diffusion-scaled process is broken down into $O(n)$ pieces of $O(n)$-long time intervals, and thereafter the diffusion-scaled process is converted to $O(n)$ pieces of fluid scaled processes. Now the properties developed for the fluid scaled process, in particular the uniform attraction property, can be applied to investigate the structure of the diffusion-scaled process and establish the state-space collapse property.

All of the methods mentioned above have appeared in the literature. To avoid the interruption to the flow of the reading, we usually do not cite the existing literature unless it is necessary for the understanding of the content. A short literature review is included in Section 11.6. The purpose of this chapter is to provide a simple exposition that illustrates the use of these methods. To this end, we choose to start with simple models and sometimes with more restrictive assumptions. Section 2 includes some elementary results so that the chapter is as self-contained as possible. Readers unfamiliar with the fluid approximation and the diffusion approximation may find it helpful to first read Chapters 5 and 6 of Chen and Yao [4]. In Section 3, we introduce the sandwich method through deriving the fluid and the diffusion approximations for a multi-server queue. In Section 4, we introduce the methods of the uniform attraction and the state-space collapse through the study of a multi-class (single-server) queue under the first-in-first-out (FIFO) service discipline. In Section 5, we study a multi-channel system, where each channel has a single server and its own queue. A single stream of arriving jobs are routed to these channels under the join-the-shortest-queue (JSQ) discipline. In deriving the diffusion approximation for this system, we illustrate the use of all of the three methods: the sandwich, the uniform attraction and the state-space collapse.

11.2 Preliminaries

Let $\mathcal{D}^K$ denote the space of the K-dimensional RCLL (Right Continuous with Left Limits) functions on $[0,\infty)$, endowed with the uniform norm. Let $\mathcal{D}_0^K = \{x \in \mathcal{D}^K : x(0) \geq 0\}$. A sequence of the processes x_n in $\mathcal{D}^K$ converges to a process x under the uniform norm is the same as that x_n converges to x on any compact set in $t \geq 0$, which we will denote by $x_n \to x$, u.o.c. ("u.o.c." is short for "uniform on any compact set"). That the sequence of the stochastic processes X_n converges to X weakly as $n \to \infty$ is denoted by $X_n \Rightarrow X$ as $n \to \infty$. Throughout the chapter, we use the Skorohod Representation Theorem (see, for example, Theorem 5.1 in Chen and Yao [4]) to convert weak convergence into almost sure convergence. So we often assume (without explicitly mentioning) that the processes have been defined on some common probability space that the convergence is almost sure converge.

Let $\{\xi_i, i \geq 1\}$ be a sequence of nonnegative i.i.d. random variables; whenever exist, we denote the mean and the standard deviation of ξ_1 by m and σ; let $\mu = 1/m$. Let

$$X(t) = \sum_{i=1}^{\lfloor t \rfloor} \xi_i \text{ and } Y(t) = \sup\{n \geq 0 : X(n) \leq t\}, \quad \text{for } t \geq 0.$$

Define the following scaled processes:

$$\begin{aligned}
\bar{X}^n(t) &= \frac{1}{n} X(nt), \\
\bar{Y}^n(t) &= \frac{1}{n} Y(nt), \\
\hat{X}^n(t) &= \sqrt{n}[\bar{X}^n(t) - mt] \equiv \frac{X(nt) - mnt}{\sqrt{n}}, \qquad \text{and} \\
\hat{Y}^n(t) &= \sqrt{n}[\bar{Y}^n(t) - \mu t] \equiv \frac{Y(nt) - \mu nt}{\sqrt{n}}.
\end{aligned}$$

We summarize some classical limit results in the following lemma

Lemma 11.2.1 (a) (Functional Strong Law of Large Numbers (FSLLN)) Suppose that ξ_1 has a finite mean $m > 0$. Then, as $n \to \infty$,

$$(\bar{X}^n, \bar{Y}^n) \to (\bar{X}, \bar{Y}) \qquad \text{u.o.c.}$$

almost surely, where $\bar{X}(t) = mt$ and $\bar{Y}(t) = \mu t$.

(b) (Functional Central Limit Theorem (FCLT)) Suppose that ξ_1 has a finite variance σ^2. Then, as $n \to \infty$,

$$(\hat{X}^n, \hat{Y}^n) \Rightarrow (\hat{X}, \hat{Y})$$

in distribution, where $\hat{X}(t) = \sigma W(t)$, $\hat{Y}(t) = -\mu \hat{X}(\mu t)$, and W is a Wiener process (i.e., a one-dimensional standard Brownian motion).

(c) (Uniform Bound for FSLLN) Suppose in addition that the variable ξ_1 has a finite $(2+\delta)$-moment for some $\delta > 0$. Let $t^* > 0$ and $u^* > 0$ be any given time lengths. Then, the following convergence holds with probability one: as $n \to \infty$,

$$\sup_{0 \le t \le nt^*} \sup_{0 \le u \le u^*} |(\bar{X}^n(t+u) - \bar{X}^n(t)) - mu| \to 0,$$
$$\sup_{0 \le t \le nt^*} \sup_{0 \le u \le u^*} |(\bar{Y}^n(t+u) - \bar{Y}^n(t)) - \mu u| \to 0.$$

The functional strong law of large numbers and the functional central limit theorem in the lemma can be found, for example, from Theorem 5.10 and 5.11 in Chen and Yao [4]; and the uniform bound on FSLLN (with a weaker assumption on the distribution of ξ_1) is from Appendix A.2 in Stolyar [17], which is based on the weak law estimate in Bramson [1].

The next two lemmas describe the reflection mapping and its least element characterization, which can be found from Chapter 6 in Chen and Yao [4] and section 2 in Chen and Shanthikumar [3], respectively.

Lemma 11.2.2 Suppose $x \in \mathcal{D}_0$. Then there is a unique pair of $y \in \mathcal{D}$ and $z \in \mathcal{D}$ such that the following relations hold for all $t \ge 0$:

$$z(t) = x(t) + y(t) \ge 0; \tag{11.1}$$
$$y(t) \text{ is non-decreasing in } t, \text{ with } y(0) = 0; \tag{11.2}$$
$$\int_0^\infty z(t) dy(t) = 0. \tag{11.3}$$

In fact, the unique pair can be written as follows,

$$y(t) = \sup_{0 \le s \le t} [-x(s)]^+,$$
$$z(t) = x(t) + \sup_{0 \le s \le t} [-x(s)]^+.$$

Denote $y = \Psi(x)$ and $z = \Phi(x)$. Then, the mappings Φ and Ψ are Lipschitz continuous on $\mathcal{D}_0$.

In the above lemma, when x is a Brownian motion, $z = \Phi(x)$ is called a *reflected* Brownian motion (RBM), and $y = \Psi(x)$ is the associated *regulator*.

Lemma 11.2.3 (a) (The Least Element Property) Suppose that $x \in \mathcal{D}_0$. If a pair of y and z satisfy the conditions (11.1) and (11.2) for all $t \ge 0$, then the following inequalities hold:

$$z(t) \ge \Phi(x)(t) \text{ and } y(t) \ge \Psi(x)(t), \quad \text{for all } t \ge 0.$$

(b) (Relaxed Dynamic Complementarity Property) Suppose that $x \in \mathcal{D}_0$. For any fixed $\varepsilon > 0$, if a pair of y and z satisfy, in addition to the condition (11.1), the following condition for all $t \geq 0$,

$$\begin{aligned} &y(t) \text{ does not increase at } t \text{ if } z(t) > \varepsilon, \text{ or equivalently} \\ &(z(t)-\varepsilon)dy(t) \leq 0. \end{aligned} \tag{11.4}$$

Then, the following inequalities hold:

$$y(t) \leq \Psi(x-\varepsilon)(t) \text{ and } z(t)-\varepsilon \leq \Phi(x-\varepsilon)(t). \tag{11.5}$$

11.3 Multi-Server Queue: Sandwich Method

Consider a queueing system with K servers, indexed by $k = 1,...,K$. The jobs arrive at the system following a counting process $A = \{A(t), t \geq 0\}$, where $A(t)$ counts the number of jobs arrived (exogenously) during $[0,t]$. Upon arrival, the job receives the service immediately if it finds at least one server being idle, otherwise, the job joins the queue with a first-come-first-served discipline. When more than one servers are available upon the arrival of a job, we assume that the job is served by the available server with the smallest index. Let $S = (S_k)$ whose kth component $S_k = \{S_k(t), t \geq 0\}$ denote the service process for server k, where $S_k(t)$ counts the number of services completed by server k during the first t units of *busy* time.

We assume that both the arrival process A and the service process S are renewal processes with their interarrival times having finite variances. By Lemma 11.2.1, we have (by invoking the Skorohod Representation Theorem) the following almost sure convergence, as $n \to \infty$,

$$\frac{1}{n}(A(n^2t) - \lambda n^2 t) \to \hat{A}(t) \text{ and } \frac{1}{n}(S(n^2t) - \mu n^2 t) \to \hat{S}(t), \text{ u.o.c. in } t \geq 0, \tag{11.1}$$

where λ is a nonnegative constant interpreted as the arrival rate, μ_k, the kth component of μ, is a constant interpreted as the service rate of server k, and $\hat{A} = \{\hat{A}(t), t \geq 0\}$ and $\hat{S} = \{\hat{S}_k(t), t \geq 0\}$ are driftless Brownian motions. The above convergence results imply (which can also be obtained directly from Lemma 11.2.1): almost surely as $n \to \infty$,

$$\frac{1}{n^2}A(n^2t) \to \lambda t \text{ and } \frac{1}{n^2}S(n^2t) \to \mu t, \text{ u.o.c. in } t \geq 0. \tag{11.2}$$

In addition, we assume that the arrival process and service processes are independent and that the heavy traffic condition holds, i.e.,

$$\lambda = \sum_{k=1}^{K} \mu_k.$$

Let $B_k(t)$ denote the total busy time of server k (i.e., the total amount of time that server k has been in service) during $[0,t]$; hence, $S_k(B_k(t))$ equals the number of service completions by server k during $[0,t]$. Then, the queue length, the number of jobs in the system, at time t, is given by the following balance equation:

$$Q(t) = Q(0) + A(t) - \sum_{k=1}^{K} S_k(B_k(t)) \geq 0. \tag{11.3}$$

In addition, the busy time process $B = \{B(t), t \geq 0\}$ with $B(t) = (B_k(t))$ and the queue length process $Q = \{Q(t), t \geq 0\}$ must satisfy the following dynamic relations:

$$0 \leq B_k(t) - B_k(s) \leq t - s \;\text{ for } t \geq s \geq 0, \qquad k = 1, ..., K, \tag{11.4}$$

$$\dot{B}_k(t) = 1 \;\text{ if } Q(t) \geq K, \qquad k = 1, ..., K. \tag{11.5}$$

(For any process $x = \{x(t), t \geq 0\}$, $\dot{x}(t)$ denotes the derivative of x at t provided the derivative exists.) The relation (11.4) has a very clear interpretation that during any time interval $[s,t]$, the total amount of busy time of server k must neither be negative nor exceed the length of the duration $(t-s)$. The relation (11.5) specifies that all of the servers must be busy when the number of jobs in the system is more than K (the number of the servers). We note that the relations (11.3)-(11.5) do not fully characterize the queue length process Q and the busy time process B. To provide a full characterization, we need to consider how an arriving job is assigned to a server when the job finds more than one servers idle. Such a complete characterization is not essential for our analysis here and hence is omitted. Interested readers can find a complete construction in the appendix of Chen and Shanthikumar [3].

As a standard procedure in the diffusion approximation, we rewrite the above relations by centering: for all $t \geq 0$,

$$Q(t) = X(t) + Y(t) \geq 0, \tag{11.6}$$

$$Y(\cdot) \text{ is non-decreasing with } Y(0) = 0, \tag{11.7}$$

$$Y(t) \text{ does not increase at time } t \text{ when } Q(t) \geq K. \tag{11.8}$$

where

$$X(t) = Q(0) + (A(t) - \lambda t) - \sum_{k=1}^{K} \left(S_k(B_k(t)) - \mu_k B_k(t)\right), \tag{11.9}$$

$$Y(t) = \sum_{k=1}^{K} \mu_k \left[t - B_k(t)\right]. \tag{11.10}$$

In rewriting (11.6), we used the heavy traffic condition, and in deriving the relation (11.8), we used the relation (11.5).

In a single server queue (where $K = 1$), the relation (11.8) is equivalent to $\int_{t=0}^{\infty} Q(t)dY(t) = 0$ and is known as the dynamic complementarity condition. Then, the relations (11.6)-(11.8) relate the processes X, Y and Q through the reflection mapping (Lemma 11.2.2), i.e., $Q = \Phi(X)$ and $Y = \Psi(X)$. In this case, since the mappings Φ and Ψ are continuous, the standard approach to establish a limit result for (the scaled version of) Q and Y is through establishing a limit result for (the corresponding scaled version of) X; the latter is usually much easier.

In the general case of the multi-server queue, the relations (11.6)-(11.8) do not uniquely characterize Y and Q for a given X. This should not be surprising since these relations are derived from the relations (11.3)-(11.5), which as we commented before do not fully characterize the queue length process Q and the busy time process B. Fortunately, by Lemma 11.2.3, we have the following lower and upper bounds: for all $t \geq 0$,

$$\Psi(X)(t) \leq Y(t) \leq \Psi(X-K)(t), \tag{11.11}$$

$$\Phi(X)(t) \leq Q(t) \leq \Phi(X-K)(t) + K. \tag{11.12}$$

When the processes X, Y and Q are taken to the fluid scale or the diffusion scale; these bounds take similar forms. In each of the fluid scale and the diffusion scale, we can show that both the upper and the lower bounds in (11.11) of the scaled Y process converge to the same limit; hence, the scaled Y process must converge to this limit. Similarly, by the inequalities in (11.12), this approach can be applied to the scaled Q process as well.[2] Therefore, the bounds in (11.11), which sandwiches the process Y, are the key in our approach in establishing the fluid approximation and the diffusion approximation for the multi-server queue.

As a first step, we establish the fluid approximation result. To this end, we introduce the following fluid scaling of the processes:

$$(\bar{Q}^n(t), \bar{X}^n(t), \bar{Y}^n(t), \bar{B}^n(t)) = \left(\frac{1}{n^2}Q(n^2t), \frac{1}{n^2}X(n^2t), \frac{1}{n^2}Y(n^2t), \frac{1}{n^2}B(n^2t)\right),$$

where the index n is a sequence of positive integers that increase to infinity. (In general, we could introduce a sequence of queueing systems, where the initial queue length, the arrival process and the service processes may vary with n. We choose not to, in order to avoid distraction from our central purpose of introducing the sandwich method.) Then,

Lemma 11.3.1 (Fluid Approximation): Suppose that the arrival process and the service processes satisfy the fluid scale convergence (11.2). Then,

$$(\bar{Q}^n, \bar{Y}^n, \bar{B}^n) \to (\bar{Q}, \bar{Y}, \bar{B}) \text{ as } n \to \infty, \text{ u.o.c.},$$

[2] On the other hand, given the convergence of the scaled X process and the scaled Y process, we can establish the convergence of the scaled Q process directly from a scaled version of the equation (11.6). In a network of multi-server queues, the corresponding inequality (11.12) does not hold, while the corresponding inequality (11.11) does hold (referring to Chen and Shanthikumar [3]). In this case, a version of the equation (11.6) is used to obtain the convergence of the scaled Q process.

where $\bar{Q}(t) = 0$, $\bar{X}(t) = 0$, $\bar{Y}(t) = 0$ and the kth component of $\bar{B}(t)$, $\bar{B}_k(t) = t$, $k = 1,...,K$, for all $t \geq 0$.

Proof. First, the process $\bar{X}^n$ can be written as

$$\bar{X}^n(t) = \frac{1}{n^2}Q(0) + \left[\frac{1}{n^2}A(n^2t) - \lambda t\right] - \sum_{k=1}^{K}\left[\frac{1}{n^2}S_k(n^2\bar{B}_k^n(t)) - \mu_k\bar{B}_k^n(t)\right] \tag{11.13}$$

then it follows from the convergence (11.2) and the fact that $0 \leq \bar{B}_k^n(t) \leq t$,

$$\bar{X}^n \to 0 \text{ as } n \to \infty, \text{ u.o.c.} \tag{11.14}$$

Next, it follows from (11.11) and (11.12) that

$$\begin{aligned}\Psi(\bar{X})^n(t) &\leq \bar{Y}^n(t) \leq \Psi(\bar{X}^n - K/n^2)(t),\\ \Phi(\bar{X})^n(t) &\leq \bar{Q}^n(t) \leq \Phi(\bar{X}^n - K/n^2)(t) + K/n^2.\end{aligned}$$

With the convergence (11.14), the above bounds imply

$$\bar{Y}^n \to \Psi(0) \equiv 0 \text{ and } \bar{Q}^n \to \Phi(0) \equiv 0 \text{ as } n \to \infty, \text{ u.o.c.}$$

Finally, it follows from (11.10),

$$\bar{Y}^n(t) = \sum_{k=1}^{K}\mu_k\left[t - \bar{B}_k^n(t)\right];$$

since $[t - \bar{B}_k^n(t)]$ is nonnegative for all $t \geq 0$ and $k = 1,...,K$, the convergence of $\bar{Y}^n$ to zero implies that $\bar{B}_k^n(t) \to t$ as $n \to \infty$, u.o.c. in $t \geq 0$, $k = 1,...,K$. □

The diffusion approximation is concerned with the following scaled processes:

$$(\hat{Q}^n(t), \hat{X}^n(t), \hat{Y}^n(t)) = \frac{1}{n}(Q(n^2t), X(n^2t), Y(n^2t)).$$

Theorem 11.3.2 (Diffusion Approximation): Suppose that the arrival process and the service processes satisfy the diffusion scale convergence (11.1). Then for almost all sample paths,

$$(\hat{Y}^n, \hat{Q}^n) \Rightarrow (\hat{Y}, \hat{Q}), \text{ as } n \to \infty, \tag{11.15}$$

where $\hat{Q} = \Phi(\hat{X})$ and $\hat{Y} = \Psi(\hat{X})$ are respectively the one-dimensional reflected Brownian motion and the associate regulator, where the Brownian motion $\hat{X} = \{\hat{X}(t), t \geq 0\}$, given by

$$\hat{X}(t) = \hat{A}(t) - \sum_{k=1}^{K}\hat{S}_k(t), \tag{11.16}$$

is a driftless Brownian motion.

Proof. First, for the diffusion scaled processes, the inequalities (11.11) and (11.12) take the form,

$$\begin{aligned}\Psi(\hat{X})^n(t) \le \hat{Y}^n(t) \le \Psi(\hat{X}^n - K/n)(t),\\ \Phi(\hat{X})^n(t) \le \hat{Q}^n(t) \le \Phi(\hat{X}^n - K/n)(t) + K/n,\end{aligned}$$

where $\hat{X}^n = \{\hat{X}^n(t), t \ge 0\}$ takes the form,

$$\hat{X}^n(t) = \frac{1}{n}Q(0) + \frac{1}{n}\left[A(n^2t) - n^2\lambda t\right] - \frac{1}{n}\sum_{k=1}^{K}\left[S_k(n^2\bar{B}_k^n(t)) - \mu_k n^2\bar{B}_k^n(t)\right].$$

It follows from the above inequalities and the continuity of the mappings Φ and Ψ (Lemma 11.2.2) that it is sufficient to show that for almost all sample paths,

$$\hat{X}^n \to \hat{X} \text{ as } n \to \infty, \text{ u.o.c.}$$

The latter convergence, with the limit given by (11.16), follows from the assumption (11.1), the random time-change theorem (cf. Chen and Yao [4], Chapter 5) and the convergence, $\bar{B}_k^n(t) \to t$ as $n \to \infty$, u.o.c. in $t \ge 0$ (which is from Lemma 11.3.1). □

11.4 A Multi-Class Queue under FIFO Service Discipline: Uniform Attraction and State-Space Collapse

Consider a single server system serving K classes of jobs. Jobs of all classes arrive exogenously, wait for service, and after service completion leave the system. Jobs are served under first-in-first-out (FIFO) discipline. Let $A = \{A(t), t \ge 0\}$ denote the arrival process, whose kth component evaluated at t, $A_k(t)$, indicates the number of arrivals of class k jobs during the time interval $[0,t]$. We assume that A_k is a renewal process whose interarrival times have a mean of $1/\lambda_k$ and variance a_k^2; the quantity λ_k is called the arrival rate of class k jobs. Let $\{v_{k,\ell}, \ell = 1,2,\ldots\}$ be a nonnegative i.i.d. sequence, where $v_{k,\ell}$ denotes the service time of the ℓth job of class k, $k = 1,2,\ldots,K$. Let $1/\mu_k$ and b_k^2 denote the mean and the variance of $v_{k,\ell}$ respectively, $k = 1,\ldots,K$. Let

$$V_k(\ell) = \sum_{\ell'=1}^{\ell} v_{k,\ell'}$$

denote the total service time of the first ℓ jobs of class k, $k = 1,\ldots,K$. For convenience, we assume that the renewal processes A_k, $k = 1,\ldots,K$, and the service time sequences $\{V_k(\ell), \ell = 1,2,\ldots\}$, $k = 1,\ldots,K$, are all mutually independent. Let $Q_k(0)$ denote the number of class k jobs initially in the system, $k = 1,...,K$, and let $W(0)$ denote the total work (measured in service time required of the server) initially in the system.

We start with a description of some performance measures of this queueing model. Let $W = \{W(t), t \geq 0\}$ be the workload process, where $W(t)$ is the total workload (measured in service time) for the server at time t, and let $Q = \{Q(t), t \geq 0\}$ be the queue length process, whose kth component evaluated at t, $Q_k(t)$, denotes the number of class k jobs in the system at time t, $k = 1, \ldots, K$. Let $B = \{B(t), t \geq 0\}$ be the K-dimensional vector busy time process, whose kth component evaluated at t, $B_k(t)$, indicates the total service time that the server has served class k jobs during $[0, t]$, $k = 1, \ldots, K$. Let

$$Y(t) = t - \sum_{k=1}^{K} B_k(t);$$

and we call $Y = \{Y(t), t \geq 0\}$ the idle time process. Let $D = \{D(t), t \geq 0\}$ be the departure process, whose kth component evaluated at t, $D_k(t)$, denote the number of class k jobs that have completed service and hence departed from the system by time t, $k = 1, \ldots, K$. Then, the queueing system must satisfy the following dynamic relations: for all time $t \geq 0$,

$$W(t) = W(0) + \sum_{k=1}^{K} V_k(A_k(t)) - t + Y(t), \tag{11.1}$$

$$Y(t) = \int_0^t 1_{\{W(s)=0\}} ds, \tag{11.2}$$

$$D_k(t + W(t)) = Q_k(0) + A_k(t), \tag{11.3}$$

$$Q_k(t) = Q_k(0) + A_k(t) - D_k(t). \tag{11.4}$$

The relation (11.1) is the work balance relation: the workload (measured in time) at time t equals the workload initially in the system plus the work arrived and subtract the work done (which is given by $[t - Y(t)]$). The relation (11.2) is the work-conserving condition: that is, the server can be idle only when there is no work in the system; hence,

$$Y(t) \text{ is non-decreasing in } t \geq 0; \text{ and } Y(0) = 0, \quad \text{and} \tag{11.5}$$

$$\int_0^\infty W(t) dY(t) = 0. \tag{11.6}$$

The relation (11.3) reflects the FIFO service discipline: for each class k, all the jobs arrived before time t (including those initially in the system) have departed the system by the time, $t + W(t)$, when all the work in the system at time t is served. The last relation (11.4) is the work balance in terms of counting the jobs arriving and departing the system.

In addition, given the FIFO service discipline, only those jobs that are initially in the system could have been served and departed from the system during the time period $[0, W(0)]$; therefore, the following condition holds:

$$\sum_{k=1}^{K} V_k(D_k(t)) \leq t < \sum_{k=1}^{K} V_k(D_k(t) + 1) \quad \text{for } t \leq W(0). \tag{11.7}$$

To see the above, recall that $V_k(D_k(t))$ is the amount of service time for the first $D_k(t)$ class-k jobs that have completed service on or before time t. Hence, the above first inequality means the total service time for all jobs that have attained service before time t must not exceed the time t, and the second inequality has a similar interpretation for any time t before all the initial jobs are served.

We close this section by introducing a sequence of systems indexed by n. Each of the networks is like the one introduced in the above, but may differ in their arrival rates and service rates (which are also indexed by n). We assume, as $n \to \infty$,

$$\lambda_k^n \to \lambda_k, \;\; \mu_k^n \to \mu_k, \quad \text{and consequently } \beta_k^n \to \beta_k, \tag{11.8}$$

where $\beta_k^n = \lambda_k^n / \mu_k^n$ and $\beta_k = \lambda_k / \mu_k$. Denote $\rho^n = \sum_{k=1}^K \beta_k^n$ and $\rho = \sum_{k=1}^K \beta_k$. Specifically, for the nth system, its arrival process of class k jobs is given by $A_k^n(t) = A_k(\lambda_k^n t/\lambda_k)$ and its service time of ℓth class k job is given by $\mu_k v_{k,\ell}/\mu_k^n$, $\ell = 1,2,\ldots$, $k = 1,\ldots,K$.

11.4.1 Fluid Approximation and Uniform Attraction

We apply the standard fluid scaling to the processes associated with the sequence of systems described above:

$$\begin{aligned}&\big(\bar{A}^n(t),\bar{B}^n(t),\bar{D}^n(t),\bar{Q}^n(t),\bar{W}^n(t),\bar{Y}^n(t)\big)\\ &= \frac{1}{n}\left(A^n(nt),B^n(nt),D^n(nt),Q^n(nt),W^n(nt),Y^n(nt)\right).\end{aligned} \tag{11.9}$$

Lemma 11.4.1 Let M be a given positive constant, and suppose $|\bar{Q}^n(0)| := \sum_{k=1}^K \bar{Q}_k^n(0) \le M$ for sufficiently large n. Then, the following conclusions hold.
(a) (Fluid limit) For any subsequence of fluid scaled processes in (11.9), there exists a further subsequence, denoted by $\mathcal{N}$, such that, along $\mathcal{N}$,

$$\big(\bar{W}^n,\bar{Q}^n,\bar{D}^n,\bar{Y}^n\big) \to \big(\bar{W},\bar{Q},\bar{D},\bar{Y}\big) \quad \text{u.o.c.} \tag{11.10}$$

for some Lipschitz continuous process $(\bar{W},\bar{Q},\bar{D},\bar{Y})$, which is referred to as the fluid limit and satisfies the following:

$$\bar{W}(t) = \bar{W}(0) + (\rho - 1)t + \bar{Y}(t) \ge 0, \quad \text{for } t \ge 0; \tag{11.11}$$

$$\bar{Y}(t) \text{ is non-decreasing in } t \ge 0, \text{ and } \bar{Y}(0) = 0; \tag{11.12}$$

$$\bar{D}_k(t + \bar{W}(t)) = \bar{Q}_k(0) + \lambda_k t, \quad \text{for } t \ge 0; \tag{11.13}$$

$$\bar{Q}_k(t) = \bar{Q}_k(0) + \lambda_k t - \bar{D}_k(t), \quad \text{for } t \ge 0; \tag{11.14}$$

$$\sum_{k=1}^K \mu_k^{-1} \bar{D}_k(t) = t, \quad \text{for } t \le \bar{W}(0). \tag{11.15}$$

(b) (Uniform attraction) Furthermore, under the heavy traffic assumption $\rho = 1$, the fluid limit satisfies the following properties:

$$\begin{aligned}
\bar{W}(t) = \bar{W}(0) \ &\leq M \sum_{k=1}^{K} \mu_k^{-1}, \ \text{ for all } t \geq 0; \\
\bar{Q}_k(t) = \lambda_k \bar{W}(t) &= \lambda_k \bar{W}(0), \ \text{ for all } t \geq \bar{W}(0).
\end{aligned} \tag{11.16}$$

Remark. Let $w = \bar{W}(0)$ denote the initial workload and define the K-dimensional vector function $Q^*(w) = \lambda w$ whose kth component, $Q_k^*(w) = \lambda_k w$. Then the property in (11.16) means that when $t \geq w$, the fluid limit of the queue length, $\bar{Q}(t) = Q^*(w) = \lambda w$. Recall that the (fluid scaled) queue length process $\bar{Q}$ is a K-dimensional process, but for large t (larger than w), it lives in a one-dimensional line (λw). As we will see in the diffusion limit (where the time is scaled by n^2 instead of n in the fluid limit), such a phenomenon happens for all time $t \geq 0$, and this property is known as the state-space collapse.

Proof. Part (a). Introduce the service renewal process $S^n = \{S^n(t), t \geq 0\}$ whose kth component evaluated at t is given by

$$S_k^n(t) = \sup\{\ell \geq 0 : V_k^n(\ell) \leq t\}.$$

First, it follows from the functional strong law of large numbers for the i.i.d. summation and the renewal process (Lemma 11.2.1) that as $n \to \infty$,

$$(\bar{A}^n(t), \bar{V}^n(t), \bar{S}^n(t)) \equiv \frac{1}{n}(A^n(nt), V^n(\lfloor nt \rfloor), S^n(nt)) \to (\lambda t, (1/\mu)t, \mu t), \tag{11.17}$$

u.o.c. in $t \geq 0$, where $1/\mu = (1/\mu_k)_{k=1}^K$. By its definition, $0 \leq \bar{B}_k^n(t) - \bar{B}_k^n(s) \leq t - s$ for all $t \geq s$ and all $n \geq 1$, $k = 1, \ldots, K$; also note that $|\bar{Q}^n(0)|$ is bounded. Hence, for any subsequence of fluid scaled processes in (11.9), there exists a further subsequence, denoted by $\mathcal{N}$, such that, along $\mathcal{N}$,

$$\bar{Q}^n(0) \to \bar{Q}(0), \quad \text{and} \quad \bar{B}^n \to \bar{B} \qquad \text{u.o.c.}, \tag{11.18}$$

for some constant vector $\bar{Q}(0)$ and Lipschitz continuous process $\bar{B} = \{\bar{B}(t), t \geq 0\}$. We show that along this same subsequence, the convergence (11.10) holds and its limit satisfies (11.11)-(11.15).

Also by their definitions,

$$\bar{Y}^n(t) = t - \sum_{k=1}^{K} \bar{B}^n(t);$$

the convergence of $\bar{B}^n$ clearly implies the convergence of $\bar{Y}^n$ along the same subsequence, and in addition, it is clear that the limit, $\bar{Y}$, of $\bar{Y}^n$ satisfies (11.12) and is Lipschitz continuous.

Letting $t = 0$ in the equation (11.3) and $t = W^n(0)$ in the inequality (11.7) (for the nth system), we derive the following inequality,

$$\sum_{k=1}^{K} V_k^n(Q_k^n(0)) \le W^n(0) < \sum_{k=1}^{K} V_k^n(Q_k^n(0)+1),$$

which in fluid scale takes the form,

$$\sum_{k=1}^{K} \bar{V}_k^n(\bar{Q}_k^n(0)) \le \bar{W}^n(0) < \sum_{k=1}^{K} \bar{V}_k^n(\bar{Q}_k^n(0)+1/n),$$

Then, the convergence of $\bar{V}^n$ in (11.17) and the convergence of $\bar{Q}_k^n(0)$ along $\mathcal{N}$ in (11.18) imply that along $\mathcal{N}$,

$$\bar{W}^n(0) \to \bar{W}(0) \equiv \sum_{k=1}^{K} \frac{1}{\mu_k} \bar{Q}_k(0)).$$

Rewriting the balance equation (11.1) for the scaled nth system, we have

$$\bar{W}^n(t) = \bar{W}^n(0) + \sum_{k=1}^{K} \bar{V}_k^n(\bar{A}_k^n(t)) - t + \bar{Y}^n(t);$$

then, in view of (11.17) and the convergence of $\bar{W}^n(0)$ and $\bar{Y}^n$, we have the convergence of $\bar{W}^n$ along $\mathcal{N}$ to the limit $\bar{W}$ as given by the equality (11.11), and the Lipschitz continuity of $\bar{W}$ follows from the same equality.

Next, by their definitions,

$$\bar{D}_k^n(t) = \bar{S}_k^n(\bar{B}_k^n(t)) \quad \text{for all } t \ge 0,$$

$k = 1, \ldots, K$. Then in view of (11.17)-(11.18), we have along the sequence $\mathcal{N}$,

$$\bar{D}_k^n(t) \to \bar{D}_k(t) \equiv \mu_k \bar{B}_k(t) \qquad \text{u.o.c. in } t \ge 0,$$

and clearly $\bar{D}_k$ is Lipschitz continuous, $k = 1, \ldots, K$. Now, rewriting a scaled version of (11.3) for the nth network, we have

$$\bar{D}_k^n(t + \bar{W}^n(t)) = \bar{Q}_k^n(0) + \bar{A}_k^n(t);$$

letting n go to infinity along the subsequence $\mathcal{N}$ obtains the relation (11.13).

Similarly, rewriting a scaled version of (11.4) for the nth network proves the convergence of $\bar{Q}^n$ along $\mathcal{N}$ and establishes (11.14); and rewriting a scaled version of (11.7) for the nth network and letting n go to infinity along $\mathcal{N}$ establishes (11.15). The equation (11.14) also establishes the Lipschitz continuity of $\bar{Q}$.

Part (b). Under the heavy traffic condition ($\rho = 1$), the equality (11.11) becomes

$$\bar{W}(t) = \bar{W}(0) + \bar{Y}(t) \ge 0;$$

this together with the condition (11.12) concludes that $\bar{W}(t) = \bar{W}(0)$ for $t \ge 0$ (which can also be seen from the reflection mapping theorem). Letting $t = \bar{W}(0)$ in (11.15) and $t = 0$ in (11.13) respectively, we have

$$\bar{W}(0) = \sum_{k=1}^{K} \mu_k^{-1} \bar{D}_k(\bar{W}(0)), \quad \text{and}$$
$$\bar{D}_k(\bar{W}(0)) = \bar{Q}_k(0).$$

The above two give the following:

$$\bar{W}(0) = \sum_{k=1}^{K} \mu_k^{-1} \bar{Q}_k(0) \le M \sum_{k=1}^{K} \mu_k^{-1}.$$

Replacing t in (11.14) by $\bar{W}(0)+t$, and then substituting (11.13) into the resulting equality, we obtain

$$\bar{Q}_k(t+\bar{W}(0)) = \lambda_k \bar{W}(0) \qquad \text{for all } t \ge 0.$$

□

11.4.2 Diffusion Approximation

As in Section 11.4.1, we consider the same sequence of the network but replacing the assumption (11.8) with the stronger assumption that as $n \to \infty$:

$$n(\rho^n - \rho) \to \theta \tag{11.19}$$

In addition, we assume that the heavy traffic condition $\rho = 1$ holds. For ease of exposition, we assume that $Q^n(0) = 0$, i.e., initially there are no jobs in the system. We note that with the specific construction of the nth network, the limit of the standard deviations for the interarrival times and service times exist: as $n \to \infty$,

$$a_k^n \to a_k \quad \text{and} \quad b_k^n \to b_k,\ k = 1, \ldots, K. \tag{11.20}$$

We apply the standard diffusion scaling (along with centering) to the key primitive and derived processes:

$$\hat{A}_k^n(t) := \frac{1}{n}\left[A_k^n(n^2 t) - \lambda_k^n n^2 t\right], \quad \hat{V}_k^n(t) = \frac{1}{n}\left[V_k^n(\lfloor n^2 t \rfloor) - (1/\mu_k^n) n^2 t\right]$$
$$\hat{Q}_k^n(t) := \frac{1}{n} Q_k^n(n^2 t), \quad \hat{Y}^n(t) := \frac{1}{n} Y^n(n^2 t), \quad \hat{W}^n(t) := \frac{1}{n} W^n(n^2 t). \tag{11.21}$$

The main theorem follows:

Theorem 11.4.2 (Diffusion Limit) Suppose that the heavy-traffic condition $\rho = 1$ holds. Then the following weak convergence holds when $n \to \infty$:

$$\left(\hat{W}^n, \hat{Y}^n, \hat{Q}^n\right) \Rightarrow \left(\hat{W}, \hat{Y}, \hat{Q}\right).$$

The limits $\hat{W} = \Phi(\hat{X})$ and $\hat{Y} = \Psi(\hat{X})$ are respectively the one-dimensional reflected Brownian motion and the associated regulator, where the Brownian motion $\hat{X}$ starts at the origin with its drift and variance respectively given by

$$\theta \qquad \text{and} \qquad \sum_{k=1}^{K} \left(\lambda^3 a^2/\mu_k^2 + \lambda_k b_k^2\right).$$

The limit $\hat{Q} = Q^*(\hat{W})$, i.e., $\hat{Q}_k(t) = \lambda_k \hat{W}(t)$ for all $t \geq 0$, $k = 1, \ldots$.

Remark. The fact that the K-dimensional queue length diffusion limit vector $\hat{Q}$ is linearly related to the one-dimensional workload diffusion limit $\hat{W}$ has been known as the state-space collapse.

Proof. First, we re-express the workload balance relation (11.1) for the nth network as follows:

$$W^n(t) = \sum_{k=1}^{K} [V^n(A_k^n(t)) - \frac{1}{\mu_k^n} A_k^n(t)] + \sum_{k=1}^{K} \frac{1}{\mu_k^n} [A_k^n(t) - \lambda_k^n t] + (\rho^n - 1)t + Y^n(t),$$

where we note $W^n(0) = 0$ is assumed. Applying the diffusion scaling to both sides of the above equation, we have

$$\hat{W}^n(t) = \hat{X}^n(t) + \hat{Y}^n(t), \tag{11.22}$$

where

$$\hat{X}^n(t) = \sum_{k=1}^{K} \hat{V}_k^n(\tilde{A}_k^n(t)) + \sum_{k=1}^{K} \frac{1}{\mu_k^n} \hat{A}_k^n(t) + n(\rho^n - 1)t; \tag{11.23}$$

and $\tilde{A}_k^n(t) := A_k^n(n^2 t)/n^2$ is a variation of the fluid-scaled process $\bar{A}_k^n$. Following (11.5) and (11.6), we also have, for each n,

$$\hat{Y}^n(t) \text{ is non-decreasing in } t \geq 0, \text{ and } \hat{Y}^n(0) = 0, \tag{11.24}$$

$$\int_0^\infty W^n(t) dY^n(t) = 0. \tag{11.25}$$

In the remaining of the proof, we adopt the standard sample path approach based on the Skorohod Representation Theorem, i.e., we assume that all of the primitive processes are defined in a probability space such that the weak convergence becomes the almost sure u.o.c. convergence. Then it follows from Lemma 11.2.1 and the random time-change theorem (see, for example, Theorem 5.3 in Chen and Yao [4]) that almost surely, as $n \to \infty$,

$$(\hat{A}^n(t), \hat{V}^n(\tilde{A}^n(t)) \to (\hat{A}(t), \hat{V}(\lambda t)), \qquad \text{u.o.c. in } t \geq 0, \tag{11.26}$$

where $\hat{A} = (A_k)$ and $\hat{V}(\lambda\cdot)$ are two independent K-dimensional driftless Brownian motions, both with independent coordinates; and the kth coordinate of $\hat{A}$, $\hat{A}_k$, has variance $\lambda_k^3 a_k^2$, and the kth coordinate of $\hat{V}(\lambda\cdot)$, $\hat{V}_k(\lambda_k\cdot)$, has variance $\lambda_k b_k^2$ respectively. The above convergence clearly implies that almost surely, as $n \longrightarrow \infty$,

$$\hat{X}^n(t) \to \hat{X}(t) \equiv \sum_{k=1}^{K} \hat{V}_k(\lambda_k t) + \sum_{k=1}^{K} \frac{1}{\mu_k}\hat{A}_k(t) + \theta t, \qquad \text{u.o.c. in } t \geq 0, (11.27)$$

where $\hat{X}$ is clearly a Brownian motion with the drift and the variance as given in the theorem.

In view of (11.22) and (11.24)-(11.25), we have $\hat{W}^n = \Phi(\hat{X}^n)$ and $\hat{Y}^n = \Psi(\hat{X}^n)$. Then, the convergence (11.27) and the continuity of the reflection mapping imply that almost surely, as $n \longrightarrow \infty$,

$$(\hat{W}^n, \hat{Y}^n) \to (\hat{W}, \hat{Y}), \qquad \text{u.o.c.},$$

where $\hat{W} = \Phi(\hat{X})$ and $\hat{Y} = \Psi(\hat{X})$.

The proof is completed by establishing a bound between Q^n and $Q^*(W^n)$, which is the key lemma for the state-space collapse, Lemma 11.4.3 below; refer to the remark immediately following the lemma. □

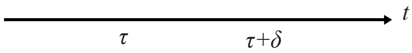
11.4.2.1 From Uniform Attraction to State-Space Collapse

Consider a fixed time interval $[\tau, \tau+\delta]$, where $\tau \geq 0$ and $\delta > 0$. Let $T > 0$ be a fixed time of a certain magnitude to be specified later. Let the index n be a large integer. Divide the time interval $[\tau, \tau+\delta]$ into a total of $\lceil n\delta/T \rceil$ segments with equal length T/n, where $\lceil\cdot\rceil$ denotes the integer ceiling. The jth segment, $j = 0, ..., \lceil n\delta/T \rceil - 1$, covers the time interval $[\tau + jT/n, \tau + (j+1)T/n]$. Note that the last interval (with j= $\lceil n\delta/T \rceil - 1$) covers a negligible piece of time beyond the right end of $[\tau, \tau+\delta]$ if $n\delta/T$ is not an integer. For notational simplicity, below we shall assume $n\delta/T$ to be an integer (i.e., omit the ceiling notation). Then, for any $t \in [\tau, \tau+\delta]$, we can write it as $t = \tau + (jT+u)/n$ for some $j = 0, \cdots, n\delta/T$ and $u \in [0, T]$. Therefore, we write

$$\hat{W}^n(t) = \hat{W}^n(\tau + \frac{jT+u}{n}) = \bar{W}^n((n\tau + jT) + u) := \bar{W}^{n,j}(u), \qquad (11.28)$$

for some real number $u \in [0, T]$ and integer $j \in [0, n\delta/T]$. That is, for each time point t, we will study the behavior of $\hat{W}^n(t)$ through the fluid scaled process, $\bar{W}^{n,j}(u)$, over the time interval $u \in [0, T]$. Similarly define $\bar{Q}_k^{n,j}(u)$ and $\bar{Y}^{n,j}(u)[= \bar{Y}^n((n\tau + jT) + u) - \bar{Y}^n(n\tau + jT)]$ as the fluid "magnifiers" of $\hat{Q}_k^n(t)$ and $\hat{Y}^n(t)$.

The above rescaling of $\hat{W}^n(t)$ is illustrated in Figure 11.1. This rescaling technique enables us to investigate the structure of diffusion-scaled processes (e.g., $\hat{W}^n(t)$) using the available results concerning the fluid-scaled processes (e.g., $\bar{W}^n(t)$).


t
τ
τ+δ

Let $\varepsilon > 0$ be any given (small) number. Then, there exists a sufficiently large T such that, for sufficiently large n, the following results hold for all non-negative integers $j < n\delta/T$:

(a) (State-space collapse)

$$|\bar{Q}^{n,j}(u) - Q^*(\bar{W}^{n,j}(u))| \leq \varepsilon, \quad \text{for all } u \in [0,T]; \tag{11.30}$$

(b)(Boundedness)

$$\bar{W}^{n,j}(u) \leq M_W := C + 2\varepsilon \text{ and } |\bar{Q}^{n,j}(u)| \leq M_Q := \sum_{k=1}^{K} \lambda_k M_W + \varepsilon, \ \text{ for all } u \in [0,T].$$

i.e., $\bar{W}^{n,j}(u)$ and $\bar{Q}^{n,j}(u)$ is uniformly bounded.

Remark. Part (a) of the lemma implies the convergence of $\hat{Q}^n$ and the state-space collapse property, $\hat{Q}(t) = Q^*(\hat{W}(t))$, in Theorem 11.4.2. To see this, we use the definition of the fluid scaling in (11.28) to rewrite the bound (11.30) as: given arbitrarily (small) number $\varepsilon > 0$, the following holds for sufficiently large index n,

$$|\hat{Q}^n(t) - Q^*(\hat{W}^n(t))| \leq \varepsilon \quad \text{for } t \in [0,\delta].$$

Part (b) of the lemma is an auxiliary result, which is required in Lemma 11.4.4 below in order to prove part (a).

Before proving Lemma 11.4.3, we present a variation of the results in Lemma 11.4.1 regarding fluid scaled processes, which will be used repeatedly.

Lemma 11.4.4 Let M be a given positive constant. Suppose $|\bar{Q}^{n,j_n}(0)| = \sum_{k=1}^{K} \bar{Q}_k^{n,j_n}(0) \leq M$ for sufficiently large n, where j_n is some integer in $[0, n\delta/T]$. Then, for any subsequence of integers $\{n\}$, there exists a further subsequence, denoted by $\mathcal{N}$, such that, along $\mathcal{N}$, the process $(\bar{Q}^{n,j_n}, \bar{W}^{n,j_n}, \bar{D}^{n,j_n}, \bar{Y}^{n,j_n})$ converge u.o.c. to the fluid limit $(\bar{Q}, \bar{W}, \bar{D}, \bar{Y})$ that satisfies all the properties described in Lemma 11.4.1.

Note that the u.o.c convergence of the primitive processes, $\bar{A}_k^{n,j_n}$ and $\bar{V}_k^{n,j_n}$ can be seen from Lemma 11.2.1 (c). Then, the proof of Lemma 11.4.4 simply replicates those of Lemma 11.4.1; hence, it is omitted.

Proof of Lemma 11.4.3. We specify the time length T (as stated in the lemma) as follows:

$$T \geq M_Q \sum_{k=1}^{K} \mu_k^{-1}.$$

Later, we will see that T is long enough so that in the fluid limit, the fluid state $\bar{Q}(t)$, starting from any initial state bounded by M_Q, will approach the fixed-point state. Below, we finish the proof in two steps.

Step 1. Here we prove the lemma for $j = 0$. Note that by way of the construction, we have

$$(\bar{W}^{n,0}(0), \bar{Q}^{n,0}(0)) = (\hat{W}^n(0), \hat{Q}^n(0)) \to (0,0), \quad \text{as } n \to \infty,$$

where we note our assumption $Q(0) = 0$. Then, from Lemma 11.4.1, we have, as $n \to \infty$,

$$(\bar{W}^{n,0}, \bar{Q}^{n,0}) \to (\bar{W}, \bar{Q}) \qquad \text{u.o.c.},$$

with $\bar{W}(u) = 0$ and $\bar{Q}(u) = 0$ for all $u \geq 0$. (Note that the convergence here is along the whole sequence of n rather than a subsequence since the limit is unique.) This immediately implies that both (a) and (b) in the lemma hold when $j = 0$ for sufficiently large n.

Step 2. We now extend the above to $j = 1, \ldots, n\delta/T$. Suppose, to the contrary, there exists a subsequence $\mathcal{N}_1$ of $\{n\}$ such that, for any $n \in \mathcal{N}_1$, at least one of the properties in (a) and (b) in the lemma do not hold for some integers $j \in [1, n\delta/T]$. Consequently, for any $n \in \mathcal{N}_1$, there exists a smallest integer, denoted as j_n, in the interval $[1, n\delta/T]$ such that at least one of the properties in (a) and (b) do not hold. To reach a contradiction, it suffices to construct an infinite subsequence $\mathcal{N}_2 \subset \mathcal{N}_1$, such that the desired properties in (a) and (b) hold for $j = j_n$ for sufficiently large $n \in \mathcal{N}_2$.

From the proof in Step 1, we assume that the properties in (a) and (b) hold for $j = 0, ..., j_n - 1$, $n \in \mathcal{N}_1$. Specifically, for $j = j_n - 1$, we have

$$|\bar{Q}^{n,j_n-1}(0)| \leq M_Q, \quad \text{for all } k \in \mathcal{N}_1.$$

Then, by Lemma 11.4.4 (cf. Lemma 11.4.1), there exists a further subsequence $\mathcal{N}_2 \subset \mathcal{N}_1$ such that

$$(\bar{W}^{n,j_n-1}, \bar{Q}^{n,j_n-1}) \to (\bar{W}, \bar{Q}) \qquad \text{u.o.c. as } n \to \infty \text{ along } \mathcal{N}_2, \tag{11.31}$$

with $|\bar{Q}(0)| \leq M_Q$. Then, we have

$$\begin{aligned} & |\bar{Q}^{n,j_n-1}(u) - Q^*(\bar{W}^{n,j_n-1}(u))| \\ \leq\ & |\bar{Q}^{n,j_n-1}(u) - \bar{Q}(u)| + |\bar{Q}(u) - Q^*(\bar{W}(u))| + |Q^*(\bar{W}(u)) - Q^*(\bar{W}^{n,j_n-1}(u))| \\ \to\ & |\bar{Q}(u) - Q^*(\bar{W}(u))| \qquad \text{u.o.c. in } u \geq 0, \text{ as } n \to \infty \text{ along } \mathcal{N}_2. \end{aligned}$$

Moreover, taking into account the choice of T and Lemma 11.4.1(b), we have

$$\bar{Q}(u) = Q^*(\bar{W}(u)) \quad \text{for all } u \geq T.$$

Therefore, for sufficiently large $n \in \mathcal{N}_2$, we have, for $u \in [0, T]$,

$$\begin{aligned} & |\bar{Q}^{n,j_n}(u) - Q^*(\bar{W}^{n,j_n}(u))| = |\bar{Q}^{n,j_n-1}(T+u) - Q^*(\bar{W}^{n,j_n-1}(T+u))| \\ & \leq |\bar{Q}(T+u) - Q^*(\bar{W}(T+u))| + \varepsilon = \varepsilon. \end{aligned} \tag{11.32}$$

That is, (a) holds with $j = j_n$ for sufficiently large $n \in \mathcal{N}_2$ ($\subset \mathcal{N}_1$).

Next, we estimate the upper bounds for $\bar{W}^{n,j_n}(u)$ and $\bar{Q}^{n,j_n}(u)$, for $u \in [0,T]$. Denote $t_1 = j_n T/n + u/n$; by definition, we have $\bar{W}^{n,j_n}(u) = \hat{W}^n(t_1)$. Let t_2 be the minimal time in $[0,t_1]$ such that $\hat{W}^n(t) > 0$ for all $t \in (t_2,t_1]$. From this definition, we know that the server will not be idle during the interval $(t_2,t_1]$, and hence

$$\hat{Y}^n(t_1) - \hat{Y}^n(t_2) = 0. \tag{11.33}$$

Observe that $\hat{W}^n(t_2) = 0$, taking into account the initial condition that $\hat{Q}^n(0) = 0$. Then, under the assumption (11.29), we estimate the upper bounds for $\bar{W}^{n,j_n}(u)$, for $u \in [0,T]$ and sufficiently large $n \in \mathcal{N}_2$, as follows,

$$\begin{aligned}\bar{W}^{n,j_n}(u) &= \hat{W}^n(t_1) = \hat{W}^n(t_2) + \big(\hat{X}^n(t_1) - \hat{X}^n(t_2)\big) + \big(\hat{Y}^n(t_1) - \hat{Y}^n(t_2)\big)\\ &= \hat{X}^n(t_1) - \hat{X}^n(t_2) \le C + \varepsilon = M_W.\end{aligned}$$

Furthermore, for the queue length process, we have

$$|\bar{Q}^{n,j_n}(u)| \le |Q^*(\bar{W}^{n,j_n}(u))| + \varepsilon = \sum_{k=1}^{K} \lambda_k \bar{W}^{n,j_n}(u) + \varepsilon \le \sum_{k=1}^{K} \lambda_k M_W + \varepsilon = M_Q.$$

The above two bounds imply that (b) holds with $j = j_n$ for sufficiently large $n \in \mathcal{N}_2$. □

11.5 Multi-Channel Queues under JSQ Routing Control

The system consists of K ($K \ge 2$) servers, indexed by $k \in \mathcal{K} := \{1,\dots,K\}$. Each server has a queue with infinite waiting room. Jobs arrive at the system following a renewal process with arrival rate λ. Upon arrival, each job is routed to one of the queues to attain service.

Let $A = \{A(t), t \ge 0\}$ denote the interarrival process, where $A(t)$ indicates the number of arrivals during the time interval $[0,t]$. We assume that A is a renewal process whose interarrival times have a mean of $1/\lambda$ and variance a^2. Let $\{v_{k,\ell}, \ell = 1,2,\dots\}$ be a nonnegative i.i.d. sequence, where $v_{k,\ell}$ denotes the time for server k to process its ℓth job, $k \in \mathcal{K}$. Let $1/\mu$ and b_k^2 denote the mean and the variance of $v_{k,\ell}$ respectively. The service rate of each server and the whole system are therefore μ and $K\mu$, respectively. Let

$$V_k(\ell) = \sum_{\ell'=1}^{\ell} v_{k,\ell'}$$

denote the total service time of the first ℓ jobs of server k, $k = 1,\dots,K$. Assume that the renewal processes A_k and the service time sequences $\{V_k(\ell), \ell = 1,2,\dots\}$, $k \in \mathcal{K}$, are all mutually independent. Let $S_k = \{S_k(t), t \ge 0\}$ denote the service process for server k, where $S_k(t)$ indicates the number of service completions by server k after serving for a total of t units of time. Clearly, the process S_k is a renewal process satisfying

$$S_k(t) = \sup\{\ell : V_k(\ell) \leq t\};$$

and its interarrival times have a mean of $1/\mu$ and variance b_k^2. The processes A and S_k are all mutually independent too.

We study the *join-the-shortest-queue* (JSQ) routing control. By JSQ, each job is routed to the shortest queue upon arrival. If there are more than one shortest queues, the tie can be broken arbitrarily; for concreteness, we assume that the job is routed to any one of the shortest queues with equal probability. Let $A_k(t)$ indicate the number of arrivals routed to server k during $[0,t]$; clearly we have,

$$\sum_{k\in\mathcal{K}} A_k(t) = A(t). \tag{11.1}$$

Let $Q_k(t)$ be the number of jobs in queue k at time t. Let $B_k(t)$ be the total amount of time that server k has served jobs by time t. We call the processes, $Q_k = \{Q_k(t), t \geq 0\}$ and $B_k = \{B_k(t), t \geq 0\}$, $k \in \mathcal{K}$, the queue length process and the busy time process respectively. Then, the following dynamic relations hold,

$$Q_k(t) = Q_k(0) + A_k(t) - S_k(B_k(t)) \geq 0, \tag{11.2}$$

$$B_k(t) = \int_0^t 1_{\{Q_k(s)>0\}} ds. \tag{11.3}$$

The first equation is a balance equation, where $Q_k(0)$ is the initial queue length, $k \in \mathcal{K}$. The second equation specifies a work-conserving condition, i.e., the server must work at its full capacity unless there are no jobs in its queue.

For convenience, we introduce the average workload process $W = \{W(t), t \geq 0\}$, the server idling process $I_k = \{I_k(t), t \geq 0\}$ and the average idling process $Y = \{Y(t), t \geq 0\}$ as follows,

$$W(t) = \frac{1}{K} \sum_{k\in\mathcal{K}} (V_k(Q_k(0) + A_k(t)) - B_k(t)), \tag{11.4}$$

$$I_k(t) = t - B_k(t) = \int_0^t 1_{\{Q_k(s)=0\}} ds, \tag{11.5}$$

$$Y(t) = \frac{1}{K} \sum_{k\in\mathcal{K}} I_k(t) = t - \frac{1}{K} \sum_{k\in\mathcal{K}} B_k(t). \tag{11.6}$$

It is easy to observe from the above expressions that

$$I_k(t),\ Y(t) \text{ are non-decreasing in } t \geq 0; \text{ and } I_k(0),\ Y(0) = 0. \tag{11.7}$$

Note that each item in the summation of (11.4) measures the workload for each individual server, and dividing it by K gives the average workload for the system (consisting K servers).

We close this section by introducing a sequence of networks indexed by n. Each of the networks is like the one introduced in the last section, but may differ in their arrival rates and mean service times (which are also indexed by n). We assume, as $n \longrightarrow \infty$,

$$\lambda^n \to \lambda, \;\; \mu^n \to \mu, \quad \text{and consequently } \rho^n \to \rho, \tag{11.8}$$

where $\rho^n = \lambda^n / K\mu^n$ and $\rho = \lambda / K\mu$. Specifically, for the nth network, its arrival process is given by $A(\lambda^n t/\lambda)$ and its service process of server k by $S_k(\mu^n t/\mu)$.

11.5.1 Fluid Approximation and Uniform Attraction

We apply the fluid scaling to the processes associated with the sequence of networks:

$$\begin{aligned} &\big(\bar{Q}^n_k(t), \bar{A}^n(t), \bar{A}^n_k(t), \bar{V}^n_k(t), \bar{S}^n_k(t), \bar{B}^n_k(t), \bar{W}^n(t), \bar{I}^n_k(t), \bar{Y}^n(t)\big) \\ &= \frac{1}{n}\left(Q^n_k(nt), A^n(nt), A^n_k(nt), V(\lfloor nt \rfloor), S^n_k(nt), B^n_k(nt), W^n(nt), I^n_k(nt), Y^n(nt)\right) \end{aligned} \tag{11.9}$$

We first establish the fluid approximation under the following assumptions: as $n \longrightarrow \infty$,

$$\bar{Q}^n_k(0) \to \bar{Q}_k(0), \qquad k \in \mathcal{K}, \tag{11.10}$$

$$\left(\bar{A}^n(t), \bar{V}^n_k(t), \bar{S}^n_k(t)\right) \to (\lambda t, \mu^{-1} t, \mu t), \quad \text{u.o.c. in } t \geq 0, \tag{11.11}$$

where $\bar{Q}_1(0) = \bar{Q}_2(0) = \cdots = \bar{Q}_K(0)$. Let $|\bar{Q}(0)| = \sum_{k \in \mathcal{K}} \bar{Q}_k(0)$. The equality, $\bar{Q}_k(0) = |\bar{Q}(0)|/K$, is to assume that the initial queue lengths at different servers are asymptotically the same (at the fluid scale). Intuitively, if the routing follows the join-the-shortest-queue, then even if the initial queue lengths are not asymptotically the same, they should become the same after some finite time; this will be established as the uniform attraction. The fluid approximation theorem follows.

Theorem 11.5.1 (FSLLN) Suppose that the sequence of the networks satisfies (11.10) and (11.11) with $\bar{Q}_k(0) = |\bar{Q}(0)|/K$, $k \in \mathcal{K}$. Then, as $n \longrightarrow \infty$,

$$\left(\bar{Q}^n_k, \bar{A}^n_k, \bar{B}^n_k\right) \to \left(\bar{Q}_k, \bar{A}_k, \bar{B}_k\right), \quad \text{u.o.c.}, \tag{11.12}$$

where, for $k = 1, \ldots, K$,

$$
\begin{aligned}
\bar{Q}_k(t) &= |\bar{Q}(t)|/K, \\
&\qquad |\bar{Q}(t)| = \left[|\bar{Q}(0)| + (\lambda - K\mu)t\right]^+; \\
\bar{A}_k(t) &= (\lambda/K)t; \\
\bar{B}_k(t) &= \begin{cases} t & t \le t_0 := \frac{|\bar{Q}(0)|}{K\mu-\lambda} \\ t_0 + \rho(t-t_0) & t > t_0 \end{cases}, \qquad \text{if } \lambda < K\mu; \\
\bar{B}_k(t) &= \rho t, \qquad \text{if } \lambda \ge K\mu.
\end{aligned}
$$

Remark: The convergence in (11.12) still holds without assuming that the initial queue lengths are asymptotically the same at all servers, i.e., $\bar{Q}_k(0) = |\bar{Q}(0)|/K$ for all $k \in \mathcal{K}$. The only added complication is in describing the limit processes.

The proof of this theorem is in the following steps. First, we show that any subsequence of the scaled processes has a further subsequence that converges u.o.c. to some Lipschitz continuous process (Lemma 11.5.2). Then we characterize the limit process in Proposition 11.5.3, which implies the characteristics of the limit in the above theorem. The fluid approximation theorem is established by showing that the limit process characterized in Proposition 11.5.3 is unique under the assumption that $\bar{Q}_1(0) = \bar{Q}_2(0) = \cdots = \bar{Q}_K(0)$.

In the theorems and lemmas in the rest of this section, we will characterize the so-called fluid limit of the derived processes under fluid scaling. These results will be used to establish the heavy traffic theorem later, and are of independent theoretical interest as well.

Lemma 11.5.2 (Fluid limit) Let M be a given positive constant, and suppose $|\bar{Q}^n(0)| = \sum_{k\in\mathcal{K}} \bar{Q}^n_k(0) \le M$ for sufficiently large n. Then, for any subsequence of fluid scaled processes in (11.9), there exists a further subsequence, denoted by $\mathcal{N}$, such that, as $n \to \infty$ along $\mathcal{N}$,

$$\left(\bar{Q}^n_k, \bar{A}^n_k, \bar{B}^n_k, \bar{W}^n, \bar{I}^n_k, \bar{Y}^n\right) \to \left(\bar{Q}_k, \bar{A}_k, \bar{B}_k, \bar{W}, \bar{I}_k, \bar{Y}\right) \quad \text{u.o.c.} \tag{11.13}$$

for some Lipschitz continuous process $(\bar{Q}_k, \bar{A}_k, \bar{B}_k, \bar{W}, \bar{I}_k, \bar{Y})$ (which will be referred to as the fluid limit). Furthermore, the fluid limit satisfies the following properties: for all $t \ge 0$,

$$\bar{Q}_k(t) = \bar{Q}_k(0) + \bar{A}_k(t) - \mu \bar{B}_k(t) \geq 0, \text{ for } k = 1, \ldots, K, \tag{11.14}$$

$$\sum_{k \in \mathcal{K}} \bar{A}_k(t) = \lambda t, \tag{11.15}$$

$$\bar{I}_k(t) = t - \bar{B}_k(t) \text{ is increasing with } \bar{I}_k(0) = 0, \text{ for } k = 1, \ldots, K, \tag{11.16}$$

$$\int_0^\infty \bar{Q}_k(t) d\bar{I}_k(t) = 0, \text{ for } k = 1, \ldots, K, \tag{11.17}$$

$$\bar{Y}(t) = \frac{1}{K} \sum_{k \in \mathcal{K}} \bar{I}_k(t) = t - \frac{1}{K} \sum_{k \in \mathcal{K}} \bar{B}_k(t), \tag{11.18}$$

$$\bar{W}(t) = \sum_{k \in \mathcal{K}} (K\mu)^{-1} \bar{Q}_k(t) = \bar{W}(0) + (\rho - 1)t + \bar{Y}(t). \tag{11.19}$$

Since the limit processes in the above theorem are all Lipschitz continuous, they are differentiable at almost all time $t \geq 0$. Below, when we write the derivative of such processes with respect to time t, we assume by default that such a time is *regular*, i.e, all the related processes are differentiable at this time t.

Proposition 11.5.3 The fluid limit in Lemma 11.5.2 satisfies, in addition to (11.14)-(11.19), the following properties:

(a)Let $t > 0$ be a regular time. If $\min_{k \in \mathcal{K}} \bar{Q}_k(t) > 0$, then,

$$\dot{\bar{Q}}_k(t) = \begin{cases} \frac{\lambda}{K^t_{\min}} - \mu & k \in \mathcal{K}^t_{\min}, \\ -\mu & k \in \mathcal{K} \setminus \mathcal{K}^t_{\min}. \end{cases}$$

If $\min_{k \in \mathcal{K}} \bar{Q}_k(t) = 0$, then,

$$\dot{\bar{Q}}_k(t) = \begin{cases} 0 & k \in \mathcal{K}^t_{\min}, \\ -\mu & k \in \mathcal{K} \setminus \mathcal{K}^t_{\min}. \end{cases}$$

Here, $\mathcal{K}^t_{\min} = \operatorname{argmin}_{k \in \mathcal{K}} \bar{Q}_k(t)$ is the set of servers with the lowest fluid level at time t, and $K^t_{\min}$ the number of queues in the set $\mathcal{K}^t_{\min}$.
Consequently, $\bar{Q}_1(t) = \cdots = \bar{Q}_K(t)$ for all $t \geq \bar{Q}(0)/\min(\lambda/K, \mu)$.

If, in addition, the heavy traffic condition ($\rho = 1$) holds, then, the following properties also hold:

(b)If $\bar{W}(0) > 0$, then $\bar{Q}_k(t) > 0$ for $t > 0$ and $k \in \mathcal{K}$.
(c)For all $t > 0$, the followings hold:

$$\bar{W}(t) = \bar{W}(0), \text{ and } \bar{B}_k(t) = t \text{ for } k \in \mathcal{K}. \tag{11.20}$$

Theorem 11.5.4 (Uniform attraction) Consider the fluid limit derived in Lemma 11.5.2, with the initial state also bounded by M. Assume that the heavy traffic condition $\rho = 1$ holds. Then, there exists a time $T_M > 0$ such that all the queues have the same length after the time T_M and the queue lengths are fixed afterward:

$$\bar{Q}_k(t) = \mu \bar{W}(t) \, (= \mu \bar{W}(0)) \quad \text{for } t \geq T_M, \; k \in \mathcal{K}. \tag{11.21}$$

The above theorem follows simply from Property 11.5.3(a,c) and the time T_M can be specified as $T_M = M/\mu$. The uniform attraction of the fluid limit is a key property used to establish the heavy traffic limit later. The attraction state (in this section, the state with all queue lengths equal) is called a fixed point in literature (e.g., Mandelbaum and Stolyar [14], Stolyar [17] and Ye and Yao [22]). For ease of presentation, we denote the fixed point state with a corresponding workload w as $Q^*(w) = \{Q^*_k(w)\}_{k\in\mathcal{K}}$, where $Q^*_k(w) = \mu w$ for $k \in \mathcal{K}$.

Remark: It is interesting to note that the above result can be generalized to the case $\rho \neq 1$. From Property 4(a), it can be seen that all queue lengths become equal within a finite time, and remain equal afterward. If $\rho < 1$, the common queue length will fall to 0, also within a finite time, and then stay unchanged at 0. If $\rho > 1$, the common queue length will then increase linearly. We leave the technical details to interested readers. Nevertheless, the above theorem, with the assumption $\rho = 1$, is sufficient for the purpose of carrying out the heavy traffic analysis in the next section.

11.5.2 Diffusion Approximation

As in Section 11.5.1, we consider the same sequence of the network but replacing the assumption (11.8) with the following stronger assumption: as $n \to \infty$:

$$n(\rho^n - \rho) \to \theta. \tag{11.22}$$

In addition, we assume that the heavy traffic condition $\rho = 1$ holds. Moreover, we also assume the existence of the limits of the standard deviations of the inter-arrival times and service times: as $n \to \infty$,

$$a^n \to a \quad \text{and} \quad b^n_k \to b_k,\ k \in \mathcal{K}. \tag{11.23}$$

Assume, for ease of exposition,

$$Q^n_k(0) = 0, \quad \text{for all } k \in \mathcal{K} \text{ and all } n \geq 1. \tag{11.24}$$

We apply the diffusion scaling (along with centering) to the associated processes:

$$\hat{A}^n(t) := \frac{1}{n}\left[A^n(n^2t) - \lambda^n n^2 t\right], \quad \hat{V}^n_k(t) := \frac{1}{n}\left[V^n_k(\lfloor n^2 t\rfloor) - \frac{1}{\mu^n}n^2 t\right],$$
$$(\hat{Q}^n_k(t), \hat{W}^n(t), \hat{Y}^n(t), \hat{I}^n_k(t)) := \frac{1}{n}(Q^n_k(n^2t), W^n(n^2t), Y^n(n^2t), I^n_k(n^2t)).$$

Theorem 11.5.5 (Diffusion Limit) Suppose the heavy-traffic condition $\rho = 1$ hold. The following weak convergence holds as $n \to \infty$:

$$\left(\hat{W}^n(t), \hat{Y}^n(t), \hat{Q}^n(t))\right) \Rightarrow \left(\hat{W}(t), \hat{Y}(t), \hat{Q}(t)\right).$$

The limits $\hat{W} = \Phi(\hat{X})$ and $\hat{Y} = \Psi(\hat{X})$ are the one-dimensional reflected Brownian motion and the regulator of the Brownian motion, where the Brownian motion $\hat{X}$ starts at the origin with its drift and variance respectively given by

$$\theta \qquad \text{and} \qquad \lambda(a^2 + \sum_{k\in\mathcal{K}} b_k^2/K^3).$$

The limit $\hat{Q} = Q^*(\hat{W})$, i.e., $\hat{Q}_k(t) = \mu\hat{W}(t)$ for all $t \geq 0$, $k \in \mathcal{K}$.

Remark: If the variances of service time, b_k^2, are uniformly bounded and the number of servers, K, is large, the variance of the limiting reflected Brownian motion in the above theorem becomes approximately λa^2. Hence, in such a situation, the system operates like a $G/D/1$ system in the limit, where the performance mainly driven by the job arrival rate, service rate and the variance of job interarrival time while the variance of service time does not matter.

Proof. First, we re-express the unscaled workload process for the nth network as follows:

$$\begin{aligned} W^n(t) = \frac{1}{K}\sum_{k\in\mathcal{K}}\left(V_k^n(A_k^n(t)) - \frac{1}{\mu^n}A_k^n(t)\right) + \frac{1}{K\mu^n}(A^n(t) - \lambda^n t) \\ +(\rho^n - 1)t + t - \frac{1}{K}\sum_{k\in\mathcal{K}} B_k^n(t), \end{aligned}$$

where we note that we assume $Q_k^n(0) = 0$. Applying the diffusion scaling to both sides of the above equation, we have

$$\hat{W}^n(t) = \hat{X}^n(t) + \hat{Y}^n(t), \tag{11.25}$$

where

$$\hat{X}^n(t) := \frac{1}{K}\sum_{k\in\mathcal{K}} \hat{V}_k^n(\tilde{A}_k^n(t)) + \frac{1}{K\mu^n}\hat{A}^n(t) + n(\rho^n - 1)t, \tag{11.26}$$

$$\hat{Y}^n(t) := n\left(t - \frac{1}{K}\sum_{k\in\mathcal{K}} \tilde{B}_k^n(t)\right); \tag{11.27}$$

and $(\tilde{A}_k^n(t), \tilde{B}_k^n(t)) := (A_k^n(n^2t)/n^2, B_k^n(n^2t)/n^2)$ is a variation of the fluid-scaled process $(\bar{A}_k^n, \bar{B}_k^n)$. Similar to (11.7), we also have, for each n,

$$\hat{Y}^n(t) \text{ is non-decreasing in } t \geq 0, \text{ and } \hat{Y}^n(0) = 0. \tag{11.28}$$

As in the proof of Theorem 11.4.2, we adopt the standard sample-path approach, and assume the following u.o.c. convergence: with probability one,

$$(\hat{A}^n, \hat{V}_k^n) \to (\hat{A}, \hat{V}_k) \quad \text{u.o.c.},$$

where $\hat{A}$ is a Brownian motion with zero mean and variance $\lambda^3 a^2$; $\hat{V}_k$ is a Brownian motion with zero mean and variance b_k^2. Below, we focus on any sample-path that satisfies the above u.o.c. convergence. Note that the results developed for the fluid limit in the last section apply to the variations like $(\tilde{A}_k^n(t), \tilde{B}_k^n(t))$ too. Then, under the JSQ routing control, we have, according to Property 11.5.3(c),

$$\tilde{A}_k^n(t) \to \bar{A}_k(t), \text{ and } \tilde{B}_k^n(t) \to \bar{B}_k(t) = t, \quad \text{u.o.c. in } t \geq 0. \tag{11.29}$$

Given the assumption $Q_k^n(0) = 0$ and thus $\bar{Q}_k(0) = 0$, it can be observed from Property 11.5.3(c) and the relation (11.14) that $\bar{A}_k(t) = \lambda t/K (= \mu t)$. Hence, under the JSQ,

$$\hat{V}_k^n(\tilde{A}_k^n(t)) \to \hat{V}_k(\lambda t/K) \text{ and } \hat{X}^n \to \hat{X} \quad \text{u.o.c. (in } t \geq 0), \tag{11.30}$$

where

$$\hat{X}(t) := \frac{1}{K} \sum_{k \in \mathcal{K}} \hat{V}_k(\lambda t/K) + \frac{1}{K\mu} \hat{A}(t) + \theta t \tag{11.31}$$

is a Brownian motion with drift and variance as given in the theorem.

Next, we apply the sandwich method to the diffusion scaled processes. From the least element characterization of the reflection mapping in Lemma 11.2.3, we have,

$$\hat{Y}^n(t) \geq \Psi(\hat{X}^n)(t) \text{ and } \hat{W}^n(t) \geq \Phi(\hat{X}^n)(t). \tag{11.32}$$

On the other hand, under JSQ, we would expect that no servers should be idle when the workload in the system is sufficiently large. Specifically, if we can show that for any $\varepsilon > 0$,

$$\begin{aligned} &\hat{Y}^n(\cdot) \text{ does not increase at } t \text{ if } \hat{W}^n(t) > \varepsilon, \text{ or} \\ &(\hat{W}^n(t) - \varepsilon) d\hat{Y}^n(t) \leq 0, \end{aligned} \tag{11.33}$$

then by Lemma 11.2.3(b), we have

$$\hat{Y}^n(t) \leq \Psi(\hat{X}^n - \varepsilon)(t) \text{ and } \hat{W}^n(t) - \varepsilon \leq \Phi(\hat{X}^n - \varepsilon)(t). \tag{11.34}$$

In view of (11.32) and (11.34), letting $n \to \infty$ and then $\varepsilon \to 0$, we have

$$(\hat{Y}^n, \hat{W}^n) \to (\hat{Y}, \hat{W}) \text{ u.o.c.} \tag{11.35}$$

To see that the relaxed complementarity condition (11.33) holds, we note that the scaled workload $\hat{W}^n(t) > \varepsilon$ (which is equivalent to $W^n(n^2 t) > n\varepsilon$) implies the workload is sufficiently large and no servers are idle for sufficiently large n. This is indeed the case, which is summarized in a key lemma, Lemma 11.5.6 in the next subsection. In that lemma (part (a)), we also establish a bound between the scaled workload process and the scaled queue length process, which allows us to conclude the convergence of $\hat{Q}^n(t) \to \hat{Q}(t) := Q^*(\hat{W}(t))$.

11.5.2.1 Complementarity and State-Space Collapse

Similar to Section 11.4.2.1, we consider a fixed time interval $[\tau, \tau+\delta]$, where $\tau \geq 0$ and $\delta > 0$; let $T > 0$ be a fixed time of a certain magnitude to be specified later; and study the diffusion scaled process $\hat{Q}_k^n$ (resp. $\hat{W}^n$, $\hat{Y}^n$ and $\hat{I}_k^n$, etc.) through the $\lceil n\delta/T \rceil$-pieces of fluid scaled processes $\bar{Q}_k^{n,j}$ (resp. $\bar{W}^{n,j}$, $\bar{Y}^{n,j}$ and $\bar{I}_k^{n,j}$, etc.), $j = 0, ..., \lceil n\delta/T \rceil - 1$.

Lemma 11.5.6 Consider the time interval $[\tau, \tau+\delta]$, with $\tau \geq 0$ and $\delta > 0$; pick a constant $C > 0$ such that

$$\sup_{t',t'' \in [\tau,\tau+\delta]} |\hat{X}(t') - \hat{X}(t'')| \leq C; \tag{11.36}$$

and suppose

$$\lim_{n\to\infty} \hat{W}^n(\tau) = \chi, \quad \text{and} \quad \lim_{n\to\infty} \hat{Q}^n(\tau) = Q^*(\chi), \tag{11.37}$$

for some constant $\chi \geq 0$. Let $\varepsilon > 0$ be any given (small) number. Then, there exists a sufficiently large T such that, for sufficiently large n, the following results hold for all non-negative integers $j < n\delta/T$:

(a)(State-space collapse)

$$|\bar{Q}^{n,j}(u) - Q^*(\bar{W}^{n,j}(u))| \leq \varepsilon, \quad \text{for all } u \in [0,T];$$

(b)(Boundedness)

$$\bar{W}^{n,j}(u) \leq \chi + C + 3\varepsilon, \quad \text{for all } u \in [0,T],$$

i.e., $\bar{W}^{n,j}(u)$ is uniformly bounded; and hence, so is $\bar{Q}^{n,j}(u)$;
(c)(Complementarity) if $\bar{W}^{n,j}(u) > \varepsilon$ for all $u \in [0,T]$, then

$$\bar{Y}^{n,j}(u) - \bar{Y}^{n,j}(0) = 0, \quad \text{for all } u \in [0,T].$$

The proof of this lemma follows the same idea as the proof for Lemma 11.4.3, though extra effort is required to establish the boundedness and complementarity properties simultaneously. Here we explain the idea behind the proof of the lemma. The detailed proof can be found in Appendix 11.7.3.

We will prove the lemma in two steps. In step 1, we show that when n is sufficiently large, the properties (a)-(c) hold for $j = 0$. First, from the condition (11.37), we know that the initial states of the processes $(\bar{W}^{n,0}(u), \bar{Q}^{n,0}(u))$ converge to a fixed point state, $(\chi, Q^*(\chi))$, as $n \to \infty$. Then, the whole processes $(\bar{W}^{n,0}(u), \bar{Q}^{n,0}(u))$ will also converge to the fixed point state within the interval $u \in [0,T]$ and stay on that state afterward, according to a fluid limit theorem (a variation of Lemma 11.5.2) and the uniform attraction theorem. Moreover, the fixed point process $Q^*(\bar{W}^{n,0}(u))$ is close to the state $Q^*(\chi)$ for $u \in [0,T]$ too, since $\bar{W}^{n,0}(u)$ is close to χ. Consequently, both processes $Q^*(\bar{W}^{n,0}(u))$ and $\bar{Q}^{n,0}(u)$ are close the $Q^*(\chi)$, which justifies the property (a) for $j=0$. As the workload $\bar{W}^{n,0}(u)$ is close to a constant χ, the boundedness property (b) becomes obvious. Finally, we note that each queue length $\bar{Q}_k^{n,0}(u)$ is close to $\mu\chi$. Thus all queue lengths can not be empty if $\chi > \varepsilon$, which implies that all servers are busy and hence the property (c).

In step 2, we extend the above to $j=1,\cdots,n\delta/T$ through induction. To this end, one may be tempted to carry out the induction in a conventional way, which we describe as follows. Assuming for sufficiently large n, the properties (a)-(c) hold for $j=0,\cdots,j_n-1$, show that they also hold for $j=j_n$ for sufficiently large n. Consider the sequence of processes $(\bar{W}^{n,j_n-1}(u),\bar{Q}^{n,j_n-1}(u))$. Since the initial states are bounded (the property (b) for j_n-1), the fluid limit theorem can be applied to show that the sequence of processes converges to a fluid limit $(\bar{W}(u),\bar{Q}(u))$. (Rigorously speaking, the convergence is along some subsequence of the network sequence.) Applying the uniform attraction theorem to the limit, we know that the fluid state $\bar{Q}(u)$ is close to the fixed point $Q^*(\bar{W}(u))$ for $u\in[T,2T]$, given that the time length T is long enough. Combining the above, we know that for sufficient large n, the process $\bar{Q}^{n,j_n-1}(u)$ is close to the fixed point $Q^*(\bar{W}^{n,j_n-1}(u))$ in the time interval $[T,2T]$. This implies the property (a) immediately, noting that the process $(\bar{W}^{n,j_n-1}(u),\bar{Q}^{n,j_n-1}(u))$, $u\in[T,2T]$, is identical to $(\bar{W}^{n,j_n(u)},\bar{Q}^{n,j_n}(u))$, $u\in[0,T]$. Similar to step 1, all the queue lengths, $\bar{Q}_k^{n,j_n}(u)$ for $k\in\mathcal{K}$, will be positive in the interval $[0,T]$ if the condition $\bar{W}^{n,j_n}(u)>\varepsilon$, $u\in[0,T]$, is satisfied, and hence the property (c) follows. Lastly, we estimate the bound for $\bar{W}^{n,j_n}(u)$ to prove the property (b). we trace the workload processes $\bar{W}^{n,j}(u)$, with the index j running backward from j_n till it hits 0 or the workload hits the level ε. Denote as j_n^0 the index at which the tracing procedure stops. Then, before the tracing stops at $j=j_n^0$, the system idling process, $\bar{Y}^{n,j}(u)$, does not vary, since the workload $\bar{W}^{n,j}(u)$ stay above ε and thus the property (c) applies. Hence, the range within which the workload processes $\bar{W}^{n,j}(u)$ vary is determined by the free processes $\bar{X}^{n,j}(u)$ for j running from j_n back to j_n^0. From the condition (11.36), the range is (roughly) bounded by C. On the other hand, when the tracing stops, the workload is close to either χ or ε. Summarizing the above, the workload $\bar{W}^{n,j_n}(u)$ is (roughly) bounded by either $\chi+C$ or $\varepsilon+C$.

Clearly, the above inductive argument does not yield a satisfactory proof, as it does not guarantee the *existence of a sufficiently large n'* such that, for all networks with larger index n ($n\geq n'$), the properties (a)-(c) hold for all $j=0,\cdots,n\delta/T$. To overcome this difficulty, the inductive argument is carried out by way of contradiction in the detailed proof; readers are referred to the appendix for details.

11.6 Notes

The readers are referred to Chen and Yao [4] and the references in their book for the fluid and the diffusion approximations to the queueing networks such as the generalized Jackson networks, the feedforward multiclass networks and some multiclass queueing networks but all with single servers. Whitt [20] provides a comprehensive reference on the stochastic-process limits.

The diffusion approximation for the multi-server queue was first obtained by Iglehart and Whitt [10, 11]. In Chen and Shanthikumar [3], the diffusion approximation is shown for the multi-server queues of generalized open and irreducible closed Jackson networks using the sandwich method. The current presentation in Section 3 follows Chen and Shanthikumar [3].

The diffusion approximation for a single server queue with multi-class jobs under a FIFO service discipline was first obtained by Peterson [15], where he established the diffusion approximation for feedforward networks of multiclass jobs under FIFO and priority service discipline. His method of the proof is different from what is presented here. That method has also been used to establish the diffusion approximation for the non-feedforward networks of multiclass queues; see, for example, Chen and Zhang [7]. The rescaling method presented in Section 4 in establishing the state-space collapse through the uniform attraction was first explicitly formulated by Bramson [1]; see also Stolyar [17, 18], Mandelbaum and Stolyar [14] and Ye and Yao [22]. The presentation of this technique here is based on Ye and Yao [22]. The state-space collapse result was probably first observed by Reiman [16]. This phenomena are exhibited in the studies of the diffusion approximation for multiclass queueing networks; e.g., Bramson [1], Bramson and Dai [2], Chen and Zhang [6, 7, 8], Chen and Ye [5], Mandelbaum and Stolyar [14], Whitt [19] and Williams [21].

The diffusion approximation for multi-channel queues to which jobs are routed based on the join-the-shortest-queue routing control was first studied in Reiman [16] and then generalized by Zhang, *et al.* [23]. The other related work includes the diffusion approximation for the flexible servers system in Mandelbaum and Stolyar [14].

11.7 Appendix

11.7.1 Proof of Lemma 11.5.2

Let $\mathcal{N}_1$ be any given subsequence of n. As the sequence of initial states $\bar{Q}^n(0)$ are bounded by the constant B, we can find a subsequence $\mathcal{N}_2$ of $\mathcal{N}_1$, such that,

$$\bar{Q}^n(0) \to \bar{Q}(0) \quad \text{as } n \to \infty \text{ along } \mathcal{N}_2. \tag{11.1}$$

As the processes $\bar{A}^n_k$ and $\bar{B}^n_k$ are RCLL and non-decreasing, we can find a further subsequence of $\mathcal{N}_2$, denoted $\mathcal{N}$, such that, as $n \to \infty$ along $\mathcal{N}$,

$$\bar{A}^n_k(t) \to \bar{A}_k(t) \quad \text{and} \quad \bar{B}^n_k(t) \to \bar{B}_k(t) \tag{11.2}$$

at any time $t \geq 0$, where the limit processes $\bar{A}_k$ and $\bar{B}_k$ are also RCLL and non-decreasing.

Now consider any time interval $[t_1, t_2]$, with $t_1 < t_2$. From the equation (11.3) (with superscript n appended properly), we have,

$$\bar{B}^n_k(t_2) - \bar{B}^n_k(t_1) = \frac{1}{n} \int_{nt_1}^{nt_2} 1_{\{Q^n_k(s)>0\}} ds \leq t_2 - t_1, \tag{11.3}$$

which implies

$$\bar{B}_k(t_2) - \bar{B}_k(t_1) \leq t_2 - t_1. \tag{11.4}$$

Hence, the process $\bar{B}_k$ is Lipschitz continuous. Next, pick any constant $c > 1$, and any time interval $[t'_1, t'_2]$ such that (a) $[t'_1, t'_2] \supset [t_1, t_2]$, (b) $t'_2 - t'_1 \leq c(t_2 - t_1)$, and (c) $\bar{A}_k(t)$ is continuous at times t'_1 and t'_2. The fact that $\bar{A}_k(t)$ is non-decreasing and therefore continuous for almost all time t ensures the existence of the times t'_1 and t'_2. Then, due to the condition (c), the convergence of $\bar{A}^n_k(t)$ holds at times t'_1 and t'_2. Therefore, we have

$$\begin{aligned} \bar{A}_k(t_2) - \bar{A}_k(t_1) &\leq \bar{A}_k(t'_2) - \bar{A}_k(t'_1) = \lim_{n\to\infty} \bar{A}^n_k(t'_2) - \bar{A}^n_k(t'_1) \\ &= \lim_{n\to\infty} \bar{A}^n(t'_2) - \bar{A}^n(t'_1) = \lambda(t'_2 - t'_1) \leq \lambda c(t_2 - t_1), \end{aligned} \tag{11.5}$$

where the first inequality is due to the non-decreasing property of the process $\bar{A}_k$, and the convergence involved is along the subsequence $\mathcal{N}$. The above implies that the process $\bar{A}_k$ is also Lipschitz continuous.

Due to the Lipschitz continuity of the limit processes $\bar{B}_k$ and $\bar{A}_k$, the convergence in (11.2) is u.o.c. of $t \geq 0$. Finally, the u.o.c. convergence of other processes in (11.13) and the Lipschitz continuity of their limits can be seen from the equations (11.2), (11.4)-(11.6) and (11.11) (also with superscript n appended properly).

The relationships in (11.14)-(11.16) and (11.18), follow simply from the relationships (11.1), (11.2), (11.5)-(11.7) and (11.11) (with the superscript n appended) by taking the limit as n goes to infinity.

To prove (11.17), it is sufficient to show that, given any interval $[t_1, t_2]$, if $\bar{Q}_k(t) > 0$ in the interval, then $\bar{I}_k(t_2) - \bar{I}_k(t_1) = 0$. Note that $\bar{Q}^n_k(t) > 0$ also holds for $t \in [t_1, t_2]$ (or $Q^n_k(t) > 0$ for $t \in [nt_1, nt_2]$) when n is sufficiently large, since $\bar{Q}^n_k$ converge to $\bar{Q}_k$ u.o.c. Therefore, we have,

$$\bar{I}^n_k(t_2) - \bar{I}^n_k(t_1) = \frac{1}{n} \int_{nt_1}^{nt_2} 1_{\{Q^n_k(s)=0\}} ds = 0.$$

Letting $n \to \infty$ yields $\bar{I}_k(t_2) - \bar{I}_k(t_1) = 0$.

Rewriting the balance equation (11.4) for the scaled nth system, we have

$$\bar{W}^n(t) = \frac{1}{K} \sum_{k\in\mathcal{K}} \bar{V}^n_k(\bar{Q}^n_k(0) + \bar{A}^n_k(t)) - t + \bar{Y}^n(t);$$

then, in view of the convergence in (11.11), the convergence of $\bar{Q}_k^n(0)$ and $\bar{Y}^n$, and the relations (11.14)-(11.15), we have the convergence of $\bar{W}^n$ along $\mathcal{N}$ to the limit $\bar{W}$ as given by (11.19).

11.7.2 Proof of Proposition 11.5.3

Prove (a). Consider the first case, $\min_{k\in\mathcal{K}}\bar{Q}_k(t) > 0$. Pick any constant $\Delta > 0$ such that $\min_{k\in\mathcal{K}\setminus\mathcal{K}_{\min}^t}\bar{Q}_k(t) - \min_{k\in\mathcal{K}}\bar{Q}_k(t) \geq \Delta$ and $\min_{k\in\mathcal{K}}\bar{Q}_k(t) \geq \Delta$. Since $\bar{Q}(t)$ is Lipschitz continuous, we can find any small time interval $[t_1,t_2]$ satisfying $0 \leq t_1 < t < t_2$, such that $\min_{k\in\mathcal{K}\setminus\mathcal{K}_{\min}^t}\bar{Q}_k(s) - \min_{k\in\mathcal{K}}\bar{Q}_k(s) \geq \Delta/2$ and $\min_{k\in\mathcal{K}}\bar{Q}_k(s) \geq \Delta/2$ for all time s in the interval. Consider the subsequence of network, also denoted as $\{n\}$ that yields the fluid limit. Since the scaled queue length processes $\bar{Q}^n$ converge (u.o.c.) to the fluid limit $\bar{Q}$ as $n \to \infty$, we have, for sufficiently large n, the following inequalities hold for all $s \in [t_1,t_2]$,

$$\min_{k\in\mathcal{K}\setminus\mathcal{K}_{\min}^t}\bar{Q}_k^n(s) - \min_{k\in\mathcal{K}}\bar{Q}_k^n(s) \geq \frac{\Delta}{4} \quad \text{and} \quad \min_{k\in\mathcal{K}}\bar{Q}_k^n(s) \geq \frac{\Delta}{4}. \tag{11.6}$$

By “un-scaling”, the first inequality above implies that, during the interval $(nt_1,nt_2]$, the shortest queue(s) should fall within the set $\mathcal{K}_{\min}^t$ and therefore all arrivals are routed to one of the queues in the set $\mathcal{K}_{\min}^t$. Hence,

$$\sum_{k\in\mathcal{K}_{\min}^t}(A_k^n(nt_2) - A_k^n(nt_1)) = A^n(nt_2) - A^n(nt_1).$$

Divided both side by n and let $n \to \infty$, the above yields

$$\sum_{k\in\mathcal{K}_{\min}^t}(\bar{A}_k(t_2) - \bar{A}_k(t_1)) = \lambda(t_2 - t_1),$$

which implies

$$\sum_{k\in\mathcal{K}_{\min}^t}\dot{\bar{A}}_k(t) = \lambda. \tag{11.7}$$

Similarly, since no job is routed to queues that are not in the set $\mathcal{K}_{\min}^t$ during the time interval $(nt_1,nt_2]$, we can show that

$$\dot{\bar{A}}_k(t) = 0 \tag{11.8}$$

for $k \in \mathcal{K}\setminus\mathcal{K}_{\min}^t$. Moreover, from the second inequality in (11.6), we see that all servers are busy during the time interval $(nt_1,nt_2]$, and therefore

$$\dot{\bar{B}}_k(t) = 1 \tag{11.9}$$

for $k \in \mathcal{K}$. From equalities (11.7) and (11.9), we have

$$\sum_{k \in \mathcal{K}_{\min}^t} \dot{\bar{Q}}_k(t) = \sum_{k \in \mathcal{K}_{\min}^t} \left(\dot{\bar{A}}_k(t) - \mu \dot{\bar{B}}_k(t) \right) = \lambda - K_{\min}^t \mu. \tag{11.10}$$

Next, we show that, for all $k \in \mathcal{K}_{\min}^t$,

$$\dot{\bar{Q}}_k(t) = \dot{\bar{Q}}^{\min}(t). \tag{11.11}$$

Here, we denote $\bar{Q}^{\min}(t) = \min_{k \in \mathcal{K}} \bar{Q}_k(t)$. Keeping in mind that $\bar{Q}_k(t) = \bar{Q}^{\min}(t)$ for $k \in \mathcal{K}_{\min}^t$, we have the followings,

$$\dot{\bar{Q}}_k(t-) = \lim_{\delta \to 0+} \frac{1}{\delta}(\bar{Q}_k(t) - \bar{Q}_k(t-\delta)) \leq \lim_{\delta \to 0+} \frac{1}{\delta}(\bar{Q}^{\min}(t) - \bar{Q}^{\min}(t-\delta)) = \dot{\bar{Q}}^{\min}(t),$$

and similarly,

$$\dot{\bar{Q}}_k(t+) = \lim_{\delta \to 0+} \frac{1}{\delta}(\bar{Q}_k(t+\delta) - \bar{Q}_k(t)) \geq \lim_{\delta \to 0+} \frac{1}{\delta}(\bar{Q}^{\min}(t+\delta) - \bar{Q}^{\min}(t)) = \dot{\bar{Q}}^{\min}(t).$$

At the given regular time t, we have $\dot{\bar{Q}}_k(t) = \dot{\bar{Q}}_k(t-) = \dot{\bar{Q}}_k(t+)$, and hence the above implies the conclusion (11.11).

Now, the equalities (11.8)-(11.11) implies the first property in (a).

Consider the second case, $\min_{k \in \mathcal{K}} \bar{Q}_k(t) = 0$. For $k \in \mathcal{K}_{\min}^t$, the queue attains the minimum length of zero, hence, $\dot{\bar{Q}}_k(t) = 0$. For $k \in \mathcal{K} \setminus \mathcal{K}_{\min}^t$, the proof follows the same lines of the first case and hence is omitted.

From the above two cases, we have, if $\bar{Q}_k(t) > \min_{k' \in \mathcal{K}} \bar{Q}_{k'}(t)$, then,

$$\frac{d}{dt}\left(\bar{Q}_k(t) - \min_{k' \in \mathcal{K}} \bar{Q}_{k'}(t) \right) \leq -\min(\lambda/K, \mu).$$

The above implies the last conclusion in property (a).

Prove (b). From the property (a) and taking into account the heavy traffic condition $\lambda = K\mu$, we have

$$\dot{\bar{Q}}^{\min}(t) \geq 0, \tag{11.12}$$

and, for some constant $\sigma > 0$,

$$\dot{\bar{Q}}^{\min}(t) \geq \sigma \quad \text{if } \mathcal{K}_{\min}^t \neq \mathcal{K}. \tag{11.13}$$

If $\bar{Q}^{\min}(0) > 0$, we have, according to the conclusion (11.12), $\bar{Q}^{\min}(t) \geq \bar{Q}^{\min}(0) > 0$ for all $t > 0$.

Suppose now $\bar{Q}^{\min}(0) = 0$. This implies that $\mathcal{K}_{\min}^0 \neq \mathcal{K}$, since $\bar{W}(0) > 0$. Hence, we have $\mathcal{K}_{\min}^t \neq \mathcal{K}$ for $t \in [0, \delta]$, where the positive number δ is chosen small enough. Then, from the conclusion in (11.13), we have

$$\dot{\bar{Q}}^{\min}(t) \geq \sigma > 0 \quad \text{for regular time } t \in [0,\delta],$$

and therefore,

$$\bar{Q}^{\min}(t) \geq \sigma t > 0 \quad \text{for } t \in [0,\delta].$$

Using the conclusion (11.12) again, we have

$$\bar{Q}^{\min}(t) \geq \bar{Q}^{\min}(\delta) \geq \sigma\delta > 0 \quad \text{for } t \geq \delta.$$

Since δ can be arbitrarily small, the above implies

$$\bar{Q}^{\min}(t) > 0 \quad \text{for } t > 0.$$

To prove that $\bar{W}(t) = \bar{W}(0)$ in the property (c), we consider two cases. Case 1, $\bar{W}(0) > 0$. Then, from the property (b), we have $\bar{Q}_k(t) > 0$ for $t > 0$ and $k \in \mathcal{K}$. From the reflection property (11.17) of the fluid limit, the above implies $\bar{B}_k(t) = t$. Then, the property that $\bar{W}(t) = \bar{W}(0)$ follows keeping in mind the heavy traffic condition $\rho = 1$.

Case 2, $\bar{W}(0) = 0$. Suppose the conclusion were not true. Then, there exists a time $t_1 > 0$ such that $\bar{W}(t_1) > 0$. Since $\bar{W}(t)$ is continuous, we can find a time $t_2 \in (0,t_1)$ such that

$$0 = \bar{W}(0) < \bar{W}(t_2) < \bar{W}(t_1).$$

However, following the argument in case 1, we can show that $\bar{W}(t) = \bar{W}(t_2)$ for all $t \geq t_2$, which implies $\bar{W}(t_1) = \bar{W}(t_2)$ and contradicts to the above inequality.

The first equality in this property (i.e., $\bar{W}(t) = \bar{W}(0)$ for all $t \geq 0$), along with the property in (11.19), implies that $\bar{Y}(t) = 0$ and hence $\bar{I}_k(t) = 0$ ($k \in \mathcal{K}$) for all $t \geq 0$. The latter is equivalent to the second equality in the property (c).

11.7.3 Proof of Lemma 11.5.6

Preparations

We first present a variation of the results in Lemma 11.5.2 regarding fluid scaled processes, which will be used repeatedly.

Lemma 11.7.1 Let M be a given positive constant. Suppose $|\bar{Q}^{n,j_n}(0)| = \sum_{k\in\mathcal{K}} \bar{Q}_k^{n,j_n}(0) \leq M$ for sufficiently large n, where j_n is some integer in $[0, n\delta/T]$. Then, for any subsequence of the processes $\left(\bar{Q}_k^{n,j_n}, \bar{A}_k^{n,j_n}, \bar{B}_k^{n,j_n}, \bar{W}^{n,j_n}, \bar{I}_k^{n,j_n}, \bar{Y}^{n,j_n}\right)$, there exists a further subsequence, denoted $\mathcal{N}$, such that, along $\mathcal{N}$, the sequence converge u.o.c. to the fluid limit $\left(\bar{Q}_k, \bar{A}_k, \bar{B}_k, \bar{W}, \bar{I}_k, \bar{Y}\right)$ that has all the properties described in Lemma 11.5.2, Proposition 11.5.3 and Theorem 11.5.4.

Note that the u.o.c convergence of the primitive processes, $\bar{A}^{n,j_n}$ and $\bar{S}_k^{n,j_n}$, the counterpart of (11.11), can be seen from Lemma 11.2.1. Then, the proof of Lemma 11.7.1

simply replicates those of Lemma 11.5.2, Proposition 11.5.3(c) and Theorem 11.5.4; hence, it is omitted.

Next, define the following constants,

$$M_{W,1} = 2\varepsilon, \ \ M_{Q,1} = K\mu M_{W,1} + \varepsilon;$$
$$M_{W,2} = \max\{M_{W,1}, \chi + \varepsilon\} + (C + \varepsilon), \ \ M_{Q,2} = K\mu M_{W,2} + \varepsilon.$$

where the numbers, ε, χ and C, are specified in the statement of the lemma under proof. The constants defined above will be used to bound processes $\hat{Q}^n(t)$ and $\hat{W}^n(t)$ for $t \in [\tau, \tau + \delta]$ and sufficiently large n. We specify the time length T (stated in the lemma under proof) as follows:

$$T \geq T_{\max\{M_{Q,1}, M_{Q,2}\}} \ , \tag{11.14}$$

where the term on the right hand side is specified in Theorem 11.5.4. Note that T is long enough so that in the fluid limit, the fluid state $\bar{Q}(t)$ will approach to the fixed-point state, from an initial state $\bar{Q}(0)$ that is bounded by $\max\{M_{Q,1}, M_{Q,2}\}$.

With the quantities specified above, we state what we want to prove, in terms of parts (b) and (c) of the lemma, in the following stronger form (part (a) remains the same): For sufficiently large n, the following results hold for all non-negative integers $j \leq n\delta/T$:

(a) $|\bar{Q}^{n,j}(u) - Q^*(\bar{W}^{n,j}(u))| \leq \varepsilon, \quad$ for all $u \in [0, T]$;
(b1)if $\bar{W}^{n,j}(u) \leq \varepsilon (< C)$ for some $u \in [0, T]$, then, for all $u \in [0, T]$,

$$\bar{W}^{n,j}(u) \leq M_{W,1}, \qquad |\bar{Q}^{n,j}(u)| \leq M_{Q,1}; \tag{11.15}$$

(b2)if $\bar{W}^{n,j}(u) > \varepsilon$ for all $u \in [0, T]$, then, for all $u \in [0, T]$,

$$\bar{W}^{n,j}(u) \leq M_{W,2} (\leq \chi + C + 3\varepsilon), \qquad |\bar{Q}^{n,j}(u)| \leq M_{Q,2}, \tag{11.16}$$

and

$$\bar{Y}^{n,j}(u) - \bar{Y}^{n,j}(0) = 0. \tag{11.17}$$

Step 1 of the Proof

Here we prove the three parts of the lemma, (a, b1, b2), for $j = 0$. Note that by way of the construction, we have

$$(\bar{W}^{n,0}(0), \bar{Q}^{n,0}(0)) = (\hat{W}^n(\tau), \hat{Q}^n(\tau)),$$

and hence,

$$(\bar{W}^{n,0}(0), \bar{Q}^{n,0}(0)) \to (\chi, Q^*(\chi)), \qquad \text{as } n \to \infty,$$

following (11.37). Then, from Lemma 11.7.1 and Theorem 11.5.4 (with $\bar{W}(0)$ and $\bar{Q}(0)$ replaced by χ and $Q^*(\chi)$ respectively), we have, as $n \to \infty$,

$$(\bar{W}^{n,0}, \bar{Q}^{n,0}) \to (\bar{W}, \bar{Q}) \qquad \text{u.o.c.},$$

with $(\bar{W}, \bar{Q})$ satisfying $(\bar{W}(u), \bar{Q}(u)) = (\chi, Q^*(\chi))$ for $u \geq 0$. (Note that the convergence here is along the whole sequence of n rather than a subsequence since the limit is unique.) Let n be sufficiently large such that

$$|\bar{W}^{n,0}(u) - \chi| \leq \min\left\{\frac{\varepsilon}{2K\mu}, \frac{\varepsilon}{4K}, \right\}, \quad \text{and} \quad |\bar{Q}^{n,0}(u) - Q^*(\chi)| \leq \min\left\{\frac{\varepsilon}{2}, \frac{\mu\varepsilon}{4}\right\}$$

for all $u \in [0, T]$. Then, we have,

$$\begin{aligned}
&|\bar{Q}^{n,0}(u) - Q^*(\bar{W}^{n,0}(u))| \leq |\bar{Q}^{n,0}(u) - Q^*(\chi)| + |Q^*(\bar{W}^{n,0}(u)) - Q^*(\chi)| \\
&\leq \min\left\{\frac{\varepsilon}{2}, \frac{\mu\varepsilon}{4}\right\} + K\mu \mid \bar{W}^{n,0}(u) - \chi \mid \leq \min\left\{\frac{\varepsilon}{2}, \frac{\mu\varepsilon}{4}\right\} + K\mu \min\left\{\frac{\varepsilon}{2K\mu}, \frac{\varepsilon}{4K}\right\} \\
&\leq \min\left\{\varepsilon, \frac{\mu\varepsilon}{2}\right\} \qquad (11.18)
\end{aligned}$$

for all $u \in [0, T]$. That is, (a) holds when $j = 0$ for sufficiently large n.

We now verify (b1, b2). First, from the established result in (a), we know that $\bar{W}^{n,0}(u)$ is arbitrarily close to χ for all $u \in [0, T]$ when n is sufficiently large. This fact directly leads to the inequalities in (b1, b2) for $j = 0$. Next, we show the complementarity in (11.17) of (b2), for $j = 0$. Note that from the conclusion in (11.18), we have

$$|\bar{Q}_k^{n,0}(u) - \mu\bar{W}^{n,0}(u)| \leq |\bar{Q}^{n,0}(u) - Q^*(\bar{W}^{n,0}(u))| \leq \frac{\mu}{2}\varepsilon.$$

and then,

$$\bar{Q}_k^{n,0}(u) \geq \mu\bar{W}^{n,0}(u) - \frac{\mu}{2}\varepsilon \geq \frac{\mu}{2}\varepsilon > 0, \qquad (11.19)$$

where the second inequality is due to the condition in property (b2). Finally, we have, for any $u \in [0, T]$,

$$\bar{I}_k^{n,0}(u) - \bar{I}_k^{n,0}(0) = \int_0^u 1_{\{\bar{Q}_k^{n,0}(u)=0\}} ds = 0,$$

where the first equality follows from the definitions of the processes $\bar{I}_k^{n,j}(u)$ and $\hat{I}_k^n(t)$, along with (11.5); and the second equality from the conclusion in (11.19). The above equality implies (11.17).

Step 2 of the Proof

We now extend the above to $j = 1, \dots, n\delta/T$. Suppose, to the contrary, there exists a subsequence $\mathcal{N}_1$ of n such that, for any $n \in \mathcal{N}_1$, at least one of the properties (a, b1, b2) does not hold for some integers $j \in [1, n\delta/T]$. Consequently, for any $n \in \mathcal{N}_1$, there exists a smallest integer, denoted as j_n, in the interval $[1, n\delta/T]$ such that at least one of the properties (a, b1, b2) does not hold. To reach a contradiction, it suffices to construct an infinite subsequence $\mathcal{N}_2' \subset \mathcal{N}_1$, such that the desired properties in (a, b1, b2) hold for $j = j_n$ for sufficiently large $n \in \mathcal{N}_2'$. To construct such a sequence, we will first find a subsequence $\mathcal{N}_2 \subset \mathcal{N}_1$ such that the property (a) holds for $j = j_n$ for sufficiently large $n \in \mathcal{N}_2$. Next, we partition $\mathcal{N}_2$ into two further subsequences, $\mathcal{N}_2 = \mathcal{N}_3 \cup \mathcal{N}_4$; and show that the conclusion of (b1) holds for sufficiently large $n \in \mathcal{N}_3' \subset \mathcal{N}_3$, and that the conclusion of (b2) holds for sufficiently large $n \in \mathcal{N}_4$. Finally, the subsequence $\mathcal{N}_2' = \mathcal{N}_3' \cup \mathcal{N}_4$ is what we need.

From the proof in Step 1, under what is assumed above, properties (a, b1, b2) hold for $j = 0, ..., j_n - 1$, $n \in \mathcal{N}_1$. Specifically, for $j = j_n - 1$, we have

$$|\bar{Q}^{n,j_n-1}(0)| \le \max\{M_{Q,1}, M_{Q,2}\}, \quad \text{for all } k \in \mathcal{N}_1.$$

Therefore, the sequence $\{\bar{Q}^{n,j_n-1}(0), n \in \mathcal{N}_1\}$ has a convergent subsequence. Then, by Lemma 11.7.1 and Lemma 11.5.2, there exists a further subsequence $\mathcal{N}_2 \subset \mathcal{N}_1$ such that

$$(\bar{W}^{n,j_n-1}, \bar{Q}^{n,j_n-1}) \to (\bar{W}, \bar{Q}) \qquad \text{u.o.c. as } n \to \infty \text{ along } \mathcal{N}_2, \tag{11.20}$$

with $|\bar{Q}(0)| \le \max\{M_{Q,1}, M_{Q,2}\}$. Then, we have

$$\begin{aligned} & |\bar{Q}^{n,j_n-1}(u) - Q^*(\bar{W}^{n,j_n-1}(u))| \\ \le\ & |\bar{Q}^{n,j_n-1}(u) - \bar{Q}(u)| + |\bar{Q}(u) - Q^*(\bar{W}(u))| + |Q^*(\bar{W}(u)) - Q^*(\bar{W}^{n,j_n-1}(u))| \\ \to\ & |\bar{Q}(u) - Q^*(\bar{W}(u))| \qquad \text{u.o.c. of } u \ge 0, \ \text{as } n \to \infty \text{ along } \mathcal{N}_2. \end{aligned}$$

Moreover, since $T \ge T_{\max\{M_{Q,1}, M_{Q,2}\}}$ and taking into account Theorem 11.5.4, we have

$$\bar{Q}(u) = Q^*(\bar{W}(u)) \quad \text{for all } u \ge T.$$

Therefore, for sufficiently large $n \in \mathcal{N}_2$, we have, for $u \in [0, T]$,

$$|\bar{Q}^{n,j_n}(u) - Q^*(\bar{W}^{n,j_n}(u))| = |\bar{Q}^{n,j_n-1}(T+u) - Q^*(\bar{W}^{n,j_n-1}(T+u))| < \varepsilon. \tag{11.21}$$

That is, (a) holds with $j = j_n$ for sufficiently large $n \in \mathcal{N}_2$ $(\subset \mathcal{N}_1)$.

Next, we partition $\mathcal{N}_2$ into $\mathcal{N}_3 \cup \mathcal{N}_4$ according to the conditions given in (b1, b2), i.e.,

$$\begin{aligned} \mathcal{N}_3 &= \{n \in \mathcal{N}_2 : \bar{W}^{n,j_n}(u) \le \varepsilon \text{ for some } u \in [0, T]\}, \\ \mathcal{N}_4 &= \{n \in \mathcal{N}_2 : \bar{W}^{n,j_n}(u) > \varepsilon \text{ for all } u \in [0, T]\}. \end{aligned}$$

Note that at least one of the two sequences $\mathcal{N}_3$ and $\mathcal{N}_4$ must be infinite.

Suppose $\mathcal{N}_3$ is infinite. Then, for each $n \in \mathcal{N}_3$, there exists a fixed $u_n \in [0,T]$ satisfying

$$\bar{W}^{n,j_n}(u_n) \leq \varepsilon. \tag{11.22}$$

Furthermore, we can choose a subsequence $\mathcal{N}_3' \subset \mathcal{N}_3$ such that, for some $u' \in [0,T]$,

$$u_n \to u' \quad \text{as } n \to \infty \text{ along } \mathcal{N}_3'.$$

Note that the convergence in (11.20) is valid for the subsequence $\mathcal{N}_3'$ $(\subset \mathcal{N}_2)$ too. Then, we have, for all $u \geq 0$,

$$\bar{W}(u) = \bar{W}(T+u') = \lim_{n\to\infty} \bar{W}^{n,j_n-1}(T+u_n) = \lim_{n\to\infty} \bar{W}^{n,j_n}(u_n) \leq \varepsilon,$$

where the first equality follows from the property (11.20) in Proposition 11.5.3; the second follows from (11.20); and the inequality follows from (11.22). Now, for sufficiently large $n \in \mathcal{N}_3'$, we have, for all $u \in [0,T]$,

$$\bar{W}^{n,j_n}(u) = \bar{W}^{n,j_n-1}(T+u) \leq \bar{W}(T+u) + \varepsilon \leq 2\varepsilon = M_{W,1} \tag{11.23}$$

$$|\bar{Q}^{n,j_n}(u)| \leq |\bar{Q}(T+u)| + \varepsilon = K\mu\bar{W}(T+u) + \varepsilon \leq K\mu M_{W,1} + \varepsilon = M_{Q,1}, \tag{11.24}$$

where the first inequality in (11.23) follows from (11.20), and so is the first inequality in (11.24). The two inequalities in (11.23) and (11.24) together imply that (b1) holds for $j = j_n$ for sufficiently large $n \in \mathcal{N}_3'$.

Next, suppose $\mathcal{N}_4$ is infinite. The convergence in (11.20) is valid for the subsequence $\mathcal{N}_4$ $(\subset \mathcal{N}_2)$ too. Similar to (11.21), we can show that for sufficiently large $n \in \mathcal{N}_4$, the following holds: for all $u \in [0,T]$,

$$|\bar{Q}^{n,j_n}(u) - Q^*(\bar{W}^{n,j_n}(u))| \leq \frac{\mu\varepsilon}{2},$$

and hence,

$$\bar{Q}_k^{n,j_n}(u) \geq Q_k^*(\bar{W}^{n,j_n}(u)) - \frac{\mu\varepsilon}{2} \geq \mu\bar{W}^{n,j_n}(u) - \frac{\mu\varepsilon}{2} \geq \frac{\mu\varepsilon}{2} > 0.$$

Similar to the argument following (11.19), the above inequality leads to the following,

$$\bar{Y}^{n,j_n}(u) - \bar{Y}^{n,j_n}(0) = 0 \quad \text{for all } u \in [0,T], \tag{11.25}$$

for sufficiently large $n \in \mathcal{N}_4$.

Using the complementarity property just established, we estimate the upper bounds for $\bar{W}^{n,j_n}(u)$ and $\bar{Q}^{n,j_n}(u)$, for $u \in [0,T]$. For a given (sufficiently large) $n \in \mathcal{N}_4$, there are two mutually exclusive cases: (i) the condition (as well as the conclusions) in (b2) holds for all $j = 0,...,j_n$; (ii) the condition in (b1) holds for some $j = 0 \leq j \leq j_n - 1$.

In the first case, the process $\bar{Y}^{n,j}(u)$ does not increase in $u \in [0,T]$, for $j = 0,...,j_n - 1$. Thus, we have, for sufficiently large $n \in \mathcal{N}_4$,

$$\begin{aligned}
\bar{W}^{n,j_n}(u) &= \bar{W}^{n,0}(0) + \sum_{j=0}^{j_n-1} \left(\bar{W}^{n,j}(T) - \bar{W}^{n,j}(0)\right) + \left(\bar{W}^{n,j_n}(u) - \bar{W}^{n,j_n}(0)\right) \\
&= \bar{W}^{n,0}(0) + \sum_{j=0}^{j_n-1} \left(\bar{X}^{n,j}(T) - \bar{X}^{n,j}(0)\right) + \left(\bar{X}^{n,j_n}(u) - \bar{X}^{n,j_n}(0)\right) \\
&\quad + \sum_{j=0}^{j_n-1} \left(\bar{Y}^{n,j}(T) - \bar{Y}^{n,j}(0)\right) + \left(\bar{Y}^{n,j_n}(u) - \bar{Y}^{n,j_n}(0)\right) \\
&= \bar{W}^{n,0}(0) + \sum_{j=0}^{j_n-1} \left(\bar{X}^{n,j}(T) - \bar{X}^{n,j}(0)\right) + \left(\bar{X}^{n,j_n}(u) - \bar{X}^{n,j_n}(0)\right) \\
&= \hat{W}^n(\tau) + \left(\hat{X}^n(\tau + j_nT/n + u/n) - \hat{X}^n(\tau)\right) \\
&\le (\chi + \varepsilon) + \left(\hat{X}(\tau + j_nT/n + u/n) - \hat{X}(\tau) + \varepsilon\right) \\
&\le (\chi + \varepsilon) + (C + \varepsilon),
\end{aligned}$$

where the first inequality follows from the convergence in (11.37) and (11.30), and the second from (11.36).

Under the case (ii), let j_n^0 be the largest integer such that the condition in (b1) holds. Thus, for all $j = j_n^0 + 1 \le j \le j_n$, the condition and results in (b2) hold, and hence $\bar{Y}^{n,j}(u)$ does not increase in $u \in [0,T]$. Then, similar to case (i), we have, for sufficiently large $n \in \mathcal{N}_4$,

$$\begin{aligned}
\bar{W}^{n,j_n}(u) &= \bar{W}^{n,j_n^0}(T) + \sum_{j=j_n^0+1}^{j_n-1} \left(\bar{W}^{n,j}(T) - \bar{W}^{n,j}(0)\right) + \left(\bar{W}^{n,j_n}(u) - \bar{W}^{n,j_n}(0)\right) \\
&= \bar{W}^{n,j_n^0}(T) + \left(\hat{X}^n(\tau + j_nT/n + u/n) - \hat{X}^n(\tau + j_n^0T/n + T/n)\right) \\
&\le M_{W,1} + (C + \varepsilon).
\end{aligned}$$

where the inequality is due to the bound (11.15) in (b1) with $j = j_n^0$ and the definition of the constant C in (11.36). Then, synthesizing the bounds in the two cases, we have, for sufficiently large $n \in \mathcal{N}_4$ and for all $u \in [0,T]$,

$$\bar{W}^{n,j_n}(u) \le \max\{(\chi + \varepsilon) + (C + \varepsilon), M_{W,1} + (C + \varepsilon)\} = M_{W,2};$$

and furthermore

$$\begin{aligned}
|\bar{Q}^{n,j_n}(u)| &\le |Q^*(\bar{W}^{n,j_n}(u))| + \varepsilon \\
&= K\mu\bar{W}^{n,j_n}(u) + \varepsilon \le K\mu M_{W,2} + \varepsilon = M_{Q,2}.
\end{aligned}$$

The above two bounds, together with the complementarity property in (11.25), imply that (b2) holds with $j = j_n$ for sufficiently large $n \in \mathcal{N}_4$.

Finally, let $\mathcal{N}_2' = \mathcal{N}_3' \cup \mathcal{N}_4$ ($\subset \mathcal{N}_2 \subset \mathcal{N}_1$). Then, the properties in (a, b1, b2) with $j = j_n$ hold for sufficiently large $n \in \mathcal{N}_2'$ ($\subset \mathcal{N}_1$). □

Acknowledgements This research is supported in part by a research grant from NSERC and grant PolyU5318/07E from Hong Kong RGC.

References

1. Bramson, M. (1998). State Space Collapse with Application to Heavy Traffic Limits for Multiclass Queueing Networks. *Queueing Systems, Theory and Applications*. **30**, 89-148.
2. Bramson, M. and J.G. Dai. (2001). Heavy traffic limits for some queueing networks. *Annals of Applied Probability*, **11**, 49-88.
3. Chen, H. and J.G. Shanthikumar. (1994). Fluid limits and diffusion approximations for networks of multi-server queues in heavy traffic, *Journal of Discrete Event Dynamic Systems*, **4**, 269-291.
4. Chen, H. and D.D. Yao. (2001). *Fundamentals of Queueing Networks: Performance, Asymptotics and Optimization*, Springer-Verlag New York, Inc.
5. Chen, H. and H.Q. Ye. (2001). Existence Condition for the Diffusion Approximations of Priority Multiclass Queueing Networks. *Queueing Systems, Theory and Applications,* **38**, 435-470.
6. Chen, H. and H. Zhang. (1996). Diffusion approximations for re-entrant lines with a first-buffer-first-served priority discipline. *Queueing Systems, Theory and Applications*, **23**, 177-195.
7. Chen, H. and H. Zhang. (2000a). Diffusion Approximations for some multiclass queueing networks with FIFO service discipline. *Mathematics of Operations Research*. **25**, 679-707.
8. Chen, H. and H. Zhang. (2000b). A sufficient condition and a necessary condition for the diffusion approximations of multiclass queueing network under priority service discipline. *Queueing Systems, Theory and Applications*. **34**, 237-268.
9. Gans, N., G. Koole and A. Mandelbaum. (2003). Telephone call centers: Tutorial, review and research prospects. *Manufacturing Service Operations Management*. **5**, 2, 79-144.
10. Iglehart, D.L. and W. Whitt. (1970a). Multiple channel queues in heavy traffic, I. *Advances in Applied Probability*. **2**, 150-177.
11. Iglehart, D.L. and W. Whitt. (1970b). Multiple channel queues in heavy traffic, II. *Advances in Applied Probability*. **2**, 355-364.
12. Itay, g. and W. Whitt. (2009). Service-level differentiation in many-server service systems via queue-ratio routing. *Operations Research* (forthcoming).
13. Kingman, J.F.C. (1965). The heavy traffic approximation in the theory of queues. In *Proceedings of Symposium on Congestion Theory*, D. Smith and W. Wilkinson (eds.), University of North Carolina Press, Chapel Hill, 137-159.
14. Mandelbaum, A. and A.L. Stolyar. (2004). Scheduling Flexible Servers with Convex Delay Costs: Heavy-Traffic Optimality of the Generalized $c\mu$-Rule. *Operations Research*, **52**, No. 6, 836-855
15. Peterson, W.P. (1991). A heavy traffic limit theorem for networks of queues with multiple customer types. *Mathematics of Operations Research*. **16**, 90-118.
16. Reiman, M.I. (1984). Some Diffusion Approximations with State Space Collapse, *Modelling and Performance Evaluation Methodology*, F. Baccelli and G. Fayolle (editors), Springer-Verlag, 209-240.
17. Stolyar, A.L. (2004). Max-Weight Scheduling in a Generalized Switch: State Space Collapse and Workload Minimization in Heavy Traffic. *Annals of Applied Probability,* **14**, 1-53.
18. A.L. Stolyar. (2005). Optimal Routing in Output-Queued Flexible Server Systems, *Probability in the Engineering and Informational Sciences*, **19**, 141-189.

19. Whitt, W. (1974). Weak Convergence Theorems for Priority Queues: Preemptive-Resume Discipline. *Journal of Applied Probability*, **8**, 74-94.
20. Whitt, W. (2002). *Stochastic-Process Limits*, Springer, New York.
21. Williams, R.J. (1998). Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. *Queueing Systems, Theory and Applications.* **30**, 27-88.
22. Ye H.Q. and David D. Yao. (2008). Heavy Traffic Optimality of Stochastic Networks under Utility-Maximizing Resource Control. *Operations Research*, **56**, No. 2, 453-470.
23. Zhang H.Q., G.H. Hsu and R. Wang. (1995). Heavy traffic limit theorems for sequence of shortest queueing systems. *Queueing Systems, Theory and Applications,* **21**, 217-238.

Chapter 12
Queueing Networks with Gaussian Inputs

Michel Mandjes

Abstract This chapter analyzes queueing systems fed by Gaussian inputs. The analysis is of an asymptotic nature, in that the number of sources is assumed large, where link bandwidth and buffer space are scaled accordingly. Relying on powerful large-deviation techniques (in particular Schilder's theorem), we identify the exponential decay rate of the overflow for the single queue. In addition we establish a number of appealing results (duality between decay rate and variance function; convexity of buffer/bandwidth trade-off curve). Then we extend the result to the tandem setting; a lower bound on the decay rate is found, which is proven to be 'tight' under specific conditions. Also approximations for the overflow probability are presented. The last part of the chapter is devoted to priority systems.

12.1 Introduction

Over the past two decades, a significant research effort has been devoted to the large-deviations analysis of queues. It has culminated in a wealth of valuable contributions to the understanding of the occurrence of rare events (such as buffer overflow) in queues. Exact computation of the overflow probability is usually a demanding task, thus motivating the search for accurate approximations and asymptotics. Large-deviations analysis usually provides a rough (logarithmic) characterization of the overflow probability (in terms of an exponential decay rate), but also insight into the system's 'path' from 'average behavior' to the rare event.

In particular, the celebrated *many-sources scaling*, introduced in a seminal paper by Weiss [25], has provided a rich framework for obtaining large-deviations results. In a many-sources setting, one considers a queueing system fed by the superposition

Michel Mandjes
Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, the Netherlands
e-mail: m.r.h.mandjes@uva.nl

R.J. Boucherie and N.M. van Dijk (eds.), *Queueing Networks: A Fundamental Approach*,
International Series in Operations Research & Management Science 154,
DOI 10.1007/978-1-4419-6472-4_12, © Springer Science+Business Media, LLC 2011

of n i.i.d. traffic sources, with the service rates and buffer thresholds scaled with n as well. When considering single queues, it is, under very mild conditions on the source behavior, possible to calculate the exponential decay of the probability $p_n(b,c)$ that the queue (fed by n sources, and emptied at a deterministic rate nc) exceeds level nb, see, e.g., [4, 5].

Beyond the single FIFO queue. Although single queues serve as a useful baseline model and provide valuable insight, they clearly have serious limitations. First of all, traffic streams usually traverse concatenations of hops (rather than just a single node). Secondly, networks increasingly support a wide variety of traffic types, with each of them having its own specific (stochastic) characteristics and Quality-of-Service requirements in terms of packet delay, loss, and throughput metrics. In order to deal with the heterogeneity in traffic types, networks will typically rely on discriminatory scheduling mechanisms to distinguish between streams of the various classes, such as priority scheduling mechanisms. Thus, a fundamental understanding of the large-deviations behavior of stochastic networks with non-FIFO scheduling is expected to play a crucial role in providing end-to-end Quality-of-Service in multi-class networks. However, only few large-deviations results are known for these more complex buffer architectures.

Gaussian traffic. As indicated above, each type of traffic has its own stochastic properties, often summarized in terms of a covariance function. One commonly distinguishes between short-range dependent input (with just a mild correlation) and long-range dependent input (in which correlations decay relatively slowly). Traditionally mainly short-range dependent models were used for analyzing the performance of communication networks. Network measurements, performed over the past decade, however, suggested that *long*-range dependent models are more appropriate [12, 22]. Evidently, ideally, one should build a theory around a class of traffic models that covers both. This explains why the Gaussian traffic model is considered to be particularly appropriate: with a specific choice of the parameters, the model corresponds to a short-range dependent process (for instance an integrated Ornstein-Uhlenbeck process), while with other choices one obtains a long-range dependent model (for instance fractional Brownian motion). It is argued in [10] that the use of Gaussian traffic models is justified as long as the aggregation is sufficiently large (both in time and number of flows), due to Central-Limit type of arguments; the limitations, in both dimensions, are further specified in [24].

Literature. A full treatment of large deviations of Gaussian queues and their applications is given in [14]. Background on large deviations can be found in [6], whereas [9] is a nice textbook on large deviations for queues. Adler [2] is a standard reference on Gaussian processes. Parts of this chapter treat material presented in [11, 13, 18].

Organization. This chapter focuses on the analysis of tandem networks fed by Gaussian inputs under the many-sources scaling. To analyze tandems, it turns out necessary to first present a number of powerful results on Gaussian processes, in particular (the generalized version of) Schilder's theorem (Section 2). To demonstrate how this (rather abstract) machinery works, we first focus in Section 3 on the single queue, and derive a number of structural results. We then shift in Section 4 our attention to tandem networks. Again, by applying Schilder's theorem, we analyze the exponential decay rate of the buffer content exceeding a predefined level; the results are bound and approximations of the decay rate, and, in a number of special cases, its exact value. Section 5 focuses on priority systems, which can be treated analogously to tandem queues. Section 6 contains a number of concluding remarks.

12.2 Preliminaries on Gaussian processes

In this section we present a number of standard results on Gaussian processes; in particular, Schilder's theorem is stated.

In general, an arrival process is an infinitely-dimensional object $(A(t), t \in \mathbb{R})$, where $A(t)$ denotes the amount of traffic generated in time interval $[0,t)$, for $t > 0$; $(A(t), t \in \mathbb{R})$ is sometimes referred to as the cumulative work process. It is noted that $A(-t)$ is to be interpreted as the negative of the amount of traffic generated in $(-t, 0]$. We also define, for $s < t$, the work arrived in time window $[s,t)$ as $A(s,t)$ (so that $A(s,t) = A(t) - A(s)$).

12.2.1 Gaussian sources

In this section we first introduce our versatile class of input processes, to which we will refer as *Gaussian sources*. For a Gaussian source, the entire probabilistic behavior of the cumulative work process can be expressed in terms of a mean traffic rate and a variance function. The mean traffic rate μ is such that $\mathbb{E}A(s,t) = \mu \cdot (t-s)$, i.e., the amount of traffic generated is proportional to the length of the interval. The variance function $v(\cdot)$ is such that $\mathbb{V}\text{ar}A(s,t) = v(t-s)$; in particular $\mathbb{V}\text{ar}A(t) = v(t)$. Let $\mathcal{N}(\mu, \sigma^2)$ denote a Normally distributed random variable with mean μ and variance σ^2.

Definition 12.2.1 — Gaussian source. $A(\cdot)$ *is a Gaussian process with stationary increments, if for all $s < t$,*

$$A(s,t) =_{\mathrm{d}} \mathcal{N}(\mu \cdot (t-s), v(t-s)).$$

We say that $A(\cdot)$ is a Gaussian source. *We call a Gaussian source* centered *if, in addition, $\mu = 0$.*

The fact that the sources introduced in Definition 12.2.1 have stationary increments, is an immediate consequence of the fact that the distribution of $A(s,t)$ just depends on the length of the time window (i.e., $t-s$), and not on its position.

The variance function $v(\cdot)$ fully determines the correlation structure of the Gaussian source. This can be seen as follows. First notice, assuming for ease $0 < s < t$, that $\mathbb{C}\mathrm{ov}(A(s),A(t)) = \mathbb{V}\mathrm{ar}A(s) + \mathbb{C}\mathrm{ov}(A(0,s),A(s,t))$. Then, using the standard property that

$$\mathbb{V}\mathrm{ar}A(0,t) = \mathbb{V}\mathrm{ar}A(0,s) + 2\,\mathbb{C}\mathrm{ov}(A(0,s),A(s,t)) + \mathbb{V}\mathrm{ar}A(s,t),$$

we find the useful relation

$$\Gamma(s,t) := \mathbb{C}\mathrm{ov}(A(s),A(t)) = \frac{1}{2}(v(t)+v(s)-v(t-s)).$$

Indeed, knowing the variance function, we can compute all covariances. In particular, the vector $(A(s_1),\ldots,A(s_d))^{\mathrm{T}}$ is distributed d-variate Normal, with mean $(\mu s_1,\ldots,\mu s_d)^{\mathrm{T}}$ and covariance matrix Σ, whose (i,j)-th entry reads

$$\Sigma_{ij} = \Gamma(s_i,s_j), \quad i,j = 1,\ldots,d.$$

The class of Gaussian sources with stationary increments is extremely rich, and this intrinsic richness is best illustrated by the multitude of possible choices for the variance function $v(\cdot)$. In fact, one could choose any function $v(\cdot)$ that gives rise to a positive semi-definite covariance function:

$$\sum_{s,t\in S} \alpha_s \mathbb{C}\mathrm{ov}(A(s),A(t))\alpha_t \geq 0,$$

for all $S \subseteq \mathbb{R}$, and $\alpha_s \in \mathbb{R}$ for all $s \in S$.

12.2.2 Classifications

We now highlight two basic classifications of Gaussian sources. These classifications can be illustrated by means of two generic types of Gaussian sources, that we also introduce in this section.

Definition 12.2.2 — fractional Brownian motion (or fBm). *A* fractional Brownian motion *source has variance function $v(\cdot)$ characterized by $v(t) = t^{2H}$, for an $H \in (0,1)$. We call H the* Hurst parameter.

The case with $H = 1/2$ is known as (ordinary) *Brownian motion*; then the increments are independent.

Definition 12.2.3 — integrated Ornstein Uhlenbeck (or iOU). *An* integrated Ornstein-Uhlenbeck *source has variance function $v(\cdot)$ characterized by $v(t) = t - 1 + e^{-t}$.*

Long-range dependence. The first way of classifying Gaussian sources relates to the correlation structure on long timescales: we are going to distinguish between short-range dependent sources and long-range dependent sources.

To this end, we first introduce the notion of correlation on timescale t, for intervals of length ε. With $t \gg \varepsilon > 0$, it is easily seen that

$$\mathbb{C}(t,\varepsilon) := \mathbb{C}\mathrm{ov}(A(0,\varepsilon),A(t,t+\varepsilon)) = \frac{1}{2}(v(t+\varepsilon)-2v(t)+v(t-\varepsilon)).$$

For ε small, and $v(\cdot)$ twice differentiable, this looks like $\varepsilon^2 v''(t)/2$. This argument shows that the 'intensity of the correlation' is expressed by the second derivative of $v(\cdot)$: 'the more convex (concave, respectively) $v(\cdot)$ at time-scale t, the stronger the positive (negative) dependence between traffic sent 'around time 0' and traffic sent 'around time t'.

The above observations can be illustrated by using the generic processes fBm and iOU. As $v''(t) = (2H)(2H-1)t^{2H-2}$, we see that for fBm the correlation is positive when $H > \frac{1}{2}$ (the higher H, the stronger this correlation; the larger t, the weaker this correlation), and negative when $H < \frac{1}{2}$ (the lower H, the stronger this correlation; the larger t, the weaker this correlation). It is readily checked that for iOU $v''(t) = e^{-t}$. In other words: the correlation is positive, and decreasing in t.

Several processes could exhibit positive correlation, but the intensity of this correlation can vary dramatically; compare the (fast!) exponential decay of $v''(t)$ for iOU traffic with the (slow!) polynomial decay of $v''(t)$ for fBm traffic. The following definition gives a classification.

Definition 12.2.4 — long-range dependence. *We call a traffic source* long-range dependent *(lrd), when the covariances* $\mathbb{C}(k,1)$ *are non-summable:*

$$\sum_{k=1}^{\infty} \mathbb{C}(k,1) = \infty,$$

and short-range dependent *(srd) when this sum is finite.*

Turning back to the case of fBm, with variance function given by $v(t) = t^{2H}$, it is easily checked that

$$\lim_{k\to\infty} \frac{\mathbb{C}(k,1)}{k^{2H-2}} = \frac{1}{2}\cdot \lim_{k\to\infty} \frac{(1+1/k)^{2H}-2+(1-1/k)^{2H}}{1/k^2} = \frac{1}{2}\cdot v''(1).$$

This entails that we have to check whether k^{2H-2} is summable or not. We conclude that Gaussian sources with this variance function are lrd iff $2H > 1$, i.e., whenever they belong to the positively correlated case. It is easily verified that, according to Definition 12.2.4, iOU is short-range dependent.

Smoothness. A second criterion to classify Gaussian processes is based on the level of smoothness of the sample paths. We coin the following definition.

Definition 12.2.5 — smoothness. *We call a Gaussian source* smooth *if, for any $t>0$,*

$$\lim_{\varepsilon\downarrow 0}\frac{\mathbb{C}\mathrm{ov}(A(0,\varepsilon),A(t,t+\varepsilon))}{\sqrt{\mathbb{V}\mathrm{ar}(A(0,\varepsilon))\mathbb{V}\mathrm{ar}(A(t,t+\varepsilon))}}=\lim_{\varepsilon\downarrow 0}\frac{\mathbb{C}(t,\varepsilon)}{v(\varepsilon)}\neq 0,$$

and non-smooth *otherwise.*

An fBm source is non-smooth, as is readily verified:

$$\lim_{\varepsilon\downarrow 0}\frac{\mathbb{C}(t,\varepsilon)}{v(\varepsilon)}=\lim_{\varepsilon\downarrow 0}\frac{1}{2}\varepsilon^{2-2H}v''(t)=0,$$

for any $t>0$ and $H\in(0,1)$. On the other hand, the iOU source is smooth, as, for any $t>0$, applying that $2v(\varepsilon)/\varepsilon^2\to 1$ as $\varepsilon\downarrow 0$,

$$\lim_{\varepsilon\downarrow 0}\frac{\mathbb{C}(t,\varepsilon)}{v(\varepsilon)}=v''(t)=e^{-t}>0.$$

Popularly speaking, one could say that Gaussian sources are smooth if there is a notion of a *traffic rate.*

12.2.3 Schilder's theorem

This subsection introduces (the generalized version of) Schilder's theorem. 'Schilder' considers the large deviations of the sample mean of Gaussian processes (i.e., infinitely-dimensional objects), as follows. Let $A_1(\cdot),A_2(\cdot),\dots$ be a sequence of i.i.d. Gaussian processes. Then consider the 'sample mean path' $n^{-1}\sum_{i=1}^n A_i(\cdot)$. For n large, it is clear that $n^{-1}\sum_{i=1}^n A_i(t)\to\mu t$, if μ is the mean rate of the Gaussian processes. 'Schilder' describes the probability of deviations from this 'mean path': it characterizes the exponential decay rate of the sample mean path $n^{-1}\sum_{i=1}^n A_i(\cdot)$ being in a remote set. Informally, a functional $\mathbb{I}(\cdot)$ is identified, such that

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n A_i(t)\approx f(t),\, t\in\mathbb{R}\right)\approx e^{-n\mathbb{I}(f)}.\tag{12.1}$$

Here $f:\mathbb{R}\to\mathbb{R}$ is a given function (or 'path'; it is a function of time); we are wondering what the probability is of $n^{-1}\sum_{i=1}^n A_i(\cdot)$ remaining 'close to' $f(\cdot)$. Evidently, $\mathbb{I}(\cdot)$ should be such that $\mathbb{I}(f_\mu)=0$, where $f_\mu(t)=\mu t$. Schilder's theorem says that, in a logarithmic sense, (12.1) is indeed true.

More formally, 'Schilder' gives an expression for the probability of the sample mean of n i.i.d. Gaussian processes (recall that this sample mean is now a *path*) being in some set $\mathbb{S}$:

$$p_n[\mathbb{S}] := \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} A_i(\cdot) \in \mathbb{S}\right) \approx \exp\left(-n \inf_{f\in\mathbb{S}} \mathbb{I}(f)\right).$$

Here the set $\mathbb{S}$ represents a collection of paths; later in this chapter we give a number of examples of such sets. An intrinsic difficulty of 'Schilder' is that $\mathbb{I}(f)$ can be given explicitly only if f is a mixture of covariance functions, as we will see below.

From the above it is concluded that, in general, finding the minimum of $\mathbb{I}(f)$ over all $f \in \mathbb{S}$ is a hard variational problem: the optimization should be done over all paths in $\mathbb{S}$ (which are infinitely-dimensional objects), and the objective function $\mathbb{I}(f)$ is only explicitly given if f is a mixture of covariance functions. However, if we succeed in finding such a minimizing $f^{\star}(\cdot)$ in $\mathbb{S}$, then this path has an appealing interpretation. Conditional on the sample-mean path being in the set $\mathbb{S}$, with overwhelming probability this happens via a path that is 'close to' $f^{\star}(\cdot)$. We call $f^{\star}$ the *most likely path* in the set $\mathbb{S}$. Put differently: the decay rate of $p_n[\mathbb{S}]$ is fully dominated by the likelihood of the most likely element in $\mathbb{S}$: as $n \to \infty$, we have that $n^{-1}\log p_n[\mathbb{S}] \to -\mathbb{I}(f^{\star})$. Knowledge of the most likely path gives often insight into the dynamics of the problem.

After having described Schilder's theorem in a heuristic manner above, we now proceed with a formal treatment of the result. It requires the introduction of a number of concepts: (i) a *path space* Ω, (ii) a *reproducing kernel Hilbert space* R, (iii) an *inner product* $\langle\cdot,\cdot\rangle_R$, and (iv) finally a *norm* $||\cdot||_R$. This norm turns out to be intimately related to the 'rate functional' $\mathbb{I}(\cdot)$. Having defined these notions, we are able to state Schilder's theorem, which we do in Theorem 12.2.7.

The framework of Schilder's theorem is formulated as follows. Consider a sequence of i.i.d. processes $A_1(\cdot), A_2(\cdot), \ldots$, distributed as a Gaussian process with variance function $v(\cdot)$. We assume for the moment that the processes are centered, but it is clear that the results for centered processes can be translated immediately into results for noncentered processes; we return to this issue in more detail in Remark 12.3.2. Define the path space Ω as

$$\Omega := \left\{\omega : \mathbb{R} \to \mathbb{R}, \text{ continuous}, \ \omega(0) = 0, \ \lim_{t\to\infty} \frac{\omega(t)}{1+|t|} = \lim_{t\to-\infty} \frac{\omega(t)}{1+|t|} = 0\right\},$$

which is a separable Banach space by imposing the norm

$$||\omega||_{\Omega} := \sup_{t\in\mathbb{R}} \frac{|\omega(t)|}{1+|t|}.$$

In [1] it is pointed out that $A_i(\cdot)$ can be realized on Ω under the following assumption, which is supposed to be in force throughout the remainder of this chapter.

Assumption 12.2.6 *There is an $\alpha < 2$ such that*

$$\lim_{t\to\infty} \frac{v(t)}{t^{\alpha}} = 0.$$

Next we introduce and define the *reproducing kernel Hilbert space* $R \subseteq \Omega$ – see [2] for a more detailed account – with the property that its elements are roughly as smooth as the covariance function $\Gamma(s,\cdot)$. We start from a 'smaller' space $R^\star$, defined by linear combinations of covariance functions:

$$R^\star := \left\{ \omega : \mathbb{R} \to \mathbb{R},\ \omega(\cdot) = \sum_{i=1}^{n} a_i \Gamma(s_i,\cdot),\ a_i, s_i \in \mathbb{R}, n \in \mathbb{N} \right\}.$$

The inner product on this space $R^\star$ is, for $\omega_a, \omega_b \in R^\star$, defined as

$$\langle \omega_a, \omega_b \rangle_R := \left\langle \sum_{i=1}^{n} a_i \Gamma(s_i,\cdot), \sum_{j=1}^{n} b_j \Gamma(s_j,\cdot) \right\rangle_R = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i b_j \Gamma(s_i,s_j); \tag{12.2}$$

notice that this implies $\langle \Gamma(s,\cdot), \Gamma(\cdot,t) \rangle_R = \Gamma(s,t)$. This inner product has the following useful property, which we refer to as the *reproducing kernel* property,

$$\omega(t) = \sum_{i=1}^{n} a_i \Gamma(s_i,t) = \left\langle \sum_{i=1}^{n} a_i \Gamma(s_i,\cdot), \Gamma(t,\cdot) \right\rangle_R = \langle \omega(\cdot), \Gamma(t,\cdot) \rangle_R. \tag{12.3}$$

From this we introduce the norm $||\omega||_R := \sqrt{\langle \omega, \omega \rangle_R}$. The closure of $R^\star$ under this norm is defined as the space R.

Having introduced the norm $||\cdot||_R$, we can now define the rate function that will apply in Schilder's theorem:

$$\mathbb{I}(\omega) := \begin{cases} \frac{1}{2}||\omega||_R^2 & \text{if } \omega \in R; \\ \infty & \text{otherwise.} \end{cases} \tag{12.4}$$

Remark that for f that can be written as a linear combination of covariance functions (i.e., $f \in R^\star$), Equations (12.2) and (12.4) yield an explicit expression for $\mathbb{I}(f)$.

Theorem 12.2.7 — (Generalized) Schilder. *Let $A_i(\cdot) \in \Omega$ be i.i.d. centered Gaussian processes, with variance function $v(\cdot)$. Then $A_1(\cdot), A_2(\cdot), \ldots$ obeys the large deviations principle with rate function $\mathbb{I}(\cdot)$, i.e.,*

(a)For any closed set $F \subset \Omega$,

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} A_i(\cdot) \in F \right) \leq - \inf_{f \in F} \mathbb{I}(f);$$

(b)For any open set $G \subset \Omega$,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} A_i(\cdot) \in G \right) \geq - \inf_{f \in G} \mathbb{I}(f).$$

Recall that this theorem, informally, states that

$$
\begin{aligned}
p_n[\mathcal{S}] &:= \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} A_i(\cdot) \in \mathcal{S}\right) \\
&\approx \exp\left(-n \inf_{f\in\mathcal{S}} \mathbb{I}(f)\right) = \exp\left(-\frac{n}{2}\inf_{f\in\mathcal{S}} ||f||_R^2\right).
\end{aligned}
$$

In other words, if we can write the probability of our interest as $p_n[\mathcal{S}]$ for some set of paths $\mathcal{S}$, then 'Schilder' provides us (at least in principle) with the corresponding decay rate.

12.3 Single queues

In this section we apply Schilder's theorem to the single queue. Before doing that, we first review Reich's theorem, describing the relation between the input process and the steady-state queue length. We then derive the logarithmic many-sources asymptotics, which enable us to establish a number of insightful structural properties (duality between decay rate and variance function; convexity of buffer/bandwidth trade-off curve).

12.3.1 Steady-state queue length

We first present a useful relation between the steady-state queue length Q and the arrival process $A(\cdot)$, which plays a central role in the remainder of this chapter. This fundamental distributional identity is often attributed to Reich [23].

Theorem 12.3.1 — Reich. *Consider an infinite-buffer queueing system, fed by an arrival process $A(\cdot,\cdot)$ with stationary increments and mean input rate μ, that served at rate C. Suppose the system is stable, i.e., $\mu < C$. Then the following distributional identity holds:*

$$Q =_{\mathrm{d}} \sup_{t\geq 0}(A(-t,0) - Ct),$$

where Q is the steady-state buffer content. If the arrival process is time-reversible, we have in addition

$$Q =_{\mathrm{d}} \sup_{t\geq 0}(A(t) - Ct).$$

Remark 12.3.2 Let $A(\cdot)$ be a Gaussian process with mean rate μ and variance function $v(\cdot)$. Consider the 'centered version' $\bar{A}(\cdot)$ of $A(\cdot)$, i.e., the Gaussian process with mean rate 0 and variance function $v(\cdot)$. With the stability condition $\mu < C$ in force, it is trivial that

$$\mathbb{P}\left(\sup_{t\geq 0}(A(t) - Ct) \geq B\right) = \mathbb{P}\left(\sup_{t\geq 0}(\bar{A}(t) - (C-\mu)t) \geq B\right).$$

As a consequence, when we have reduced the service rate C by the mean rate μ of the input process, we can restrict ourselves, without loss of generality, to considering just centered sources. ◇

12.3.2 Logarithmic asymptotics

We now study the logarithmic asymptotics of the probability that the buffer content under the many-sources scaling, defined as Q_n, exceeds nb: applying 'Reich',

$$\begin{aligned} p_n(b,c) &:= \mathbb{P}(Q_n \geq nb) = \mathbb{P}\left(\sup_{t\geq 0}\left(\sum_{i=1}^{n} A_i(-t,0) - nct\right) \geq nb\right) \\ &= \mathbb{P}\left(\sup_{t\geq 0}\left(\frac{1}{n}\sum_{i=1}^{n} A_i(-t,0) - ct\right) \geq b\right). \end{aligned}$$

To use 'Schilder' we have to define a set of 'overflow paths':

$$\mathcal{S}^{(\mathrm{f})} := \{f \in \Omega : \exists t \geq 0 : -f(-t) \geq b + ct\}.$$

Here we use the superscript (f) as a mnemonic for FIFO, since we consider single work-conserving queues here, of which the FIFO queue is the most prominent example. Clearly, the observation that $A(-t,0) \equiv -A(-t)$ shows that indeed $p_n(b,c) = p_n[\mathcal{S}^{(\mathrm{f})}]$. This entails that we can apply Schilder's theorem to obtain

$$\lim_{n\to\infty} \frac{1}{n} \log p_n(b,c) = -\inf_{f\in\mathcal{S}^{(\mathrm{f})}} \mathbb{I}(f).$$

We obtain the following result [1]; with Remark 12.3.2 in mind, we restrict ourselves without loss of generality to centered sources.

Theorem 12.3.3 — Logarithmic asymptotics. *For any $b,c > 0$,*

$$I_c^{(\mathrm{f})}(b) := -\lim_{n\to\infty} \frac{1}{n} \log p_n(b,c) = \inf_{t\geq 0} \frac{(b+ct)^2}{2v(t)}. \tag{12.1}$$

Proof. Define $\mathcal{S}_t^{(\mathrm{f})} := \{f \in \Omega : -f(-t) \geq b + ct\}$, so that $\mathcal{S}^{(\mathrm{f})}$ is the union over the $\mathcal{S}_t^{(\mathrm{f})}$. Observe that

$$p_n[\mathcal{S}_t^{(\mathrm{f})}] = \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} A_i(-t,0) \geq a + \nu t\right).$$

Cramér's theorem [6] now entails that

$$\lim_{n\to\infty}\frac{1}{n}\log p_n[\mathbb{S}_t^{(\mathrm{f})}] = -\sup_\theta\left(\theta(b+ct) - \log\mathbb{E}e^{\theta A(-t,0)}\right) = \frac{(b+ct)^2}{2v(t)}.$$

Using that $\mathbb{S}^{(\mathrm{f})} = \cup_{t\geq 0}\mathbb{S}_t^{(\mathrm{f})}$, application of 'Schilder' yields that

$$\lim_{n\to\infty}\frac{1}{n}\log p_n[\mathbb{S}^{(\mathrm{f})}] = -\inf_{f\in\mathbb{S}^{(\mathrm{f})}}\mathbb{I}(f) = -\inf_{t\geq 0}\left(\inf_{f\in\mathbb{S}_t^{(\mathrm{f})}}\mathbb{I}(f)\right).$$

This implies the stated. □

We call the decay rate $I_c^{(\mathrm{f})}(b)$, seen as a function of the buffer size b, and with c held fixed, the *loss curve*. In this chapter the impact of b on the optimizing t in (12.1) plays a crucial rôle; we therefore use the notation $t(b)$. The path $f^\star \in \mathbb{S}^{(\mathrm{f})}$ that optimizes $\mathbb{I}(f)$ is

$$\begin{aligned} f^\star(r) &:= \mathbb{E}(A(r) \mid A(-t(b),0) = b + ct(b)) \\ &= \frac{\Gamma(r,-t(b))}{v(t(b))}(b+ct(b)); \end{aligned} \tag{12.2}$$

we call this the *most likely path to overflow.* Here $-t(b)$ can be interpreted as the most likely time at which the buffer starts to build up in order to exceed level nb at time 0; we therefore call $t(b)$ the *most likely timescale of overflow.*

Example 12.3.4 We consider a Gaussian queue with fBm input, and show that the loss curve is concave for $H > \frac{1}{2}$, and convex for $H < \frac{1}{2}$.

Take $v(t) = t^{2H}$, for some $H \in (0,1)$. If we perform the optimization in the right-hand side of (12.1), we obtain, for $b > 0$,

$$t(b) = \frac{b}{c}\frac{H}{1-H}, \quad I_c^{(\mathrm{f})}(b) = \frac{1}{2}\left(\frac{b}{1-H}\right)^{2-2H}\left(\frac{c}{H}\right)^{2H}. \tag{12.3}$$

We see that the loss curve $I_c^{(\mathrm{f})}(\cdot)$ is convex (concave) when the Hurst parameter is smaller (larger) than $\frac{1}{2}$. ◇

12.3.3 The shape of the loss curve

We now study the relationship between the correlation structure of the sources and the shape of the curve $I_c^{(\mathrm{f})}(\cdot)$. Our main result is that there is a neat connection between positive (negative) correlations and concavity (convexity) of the loss curve. First we describe, on an intuitive level, what convexity and concavity of the loss curve mean. Evidently, $I_c^{(\mathrm{f})}(\cdot)$ is increasing. It is important to notice that, clearly, the steeper $I_c^{(\mathrm{f})}(\cdot)$ at some buffer size b, the higher the marginal benefits of an additional unit of buffering (where 'benefits' are in terms of reducing the overflow probability).

If $I_c^{(\mathrm{f})}(\cdot)$ is *convex*, then adding buffering capacity is getting more and more beneficial; if $I_c^{(\mathrm{f})}(\cdot)$ is concave, then the benefit of buffering becomes smaller and smaller. This motivates the examination of the characteristics of the shape of the loss curve $I_c^{(\mathrm{f})}(\cdot)$.

In more detail, the main result of this subsection is that we show that the curve $I_c^{(\mathrm{f})}(\cdot)$ is convex (concave) in b, *if and only if* the Gaussian input exhibits negative (positive) correlations on the time-scale $t(b)$ (defined earlier in this section). All proofs are elementary, and add insight into the marginal benefits of buffering, i.e., the nature of $I_c^{(\mathrm{f})}(\cdot)$ (in terms of its derivative and second derivative with respect to the buffer size b). We impose the following (mild) technical assumption on $v(\cdot)$ that guarantees uniqueness of $t^\star(b)$ for all b. Define the standard deviation function by $\varsigma(t) := \sqrt{v(t)}$.

Assumption 12.3.5 *The following two assumptions are imposed on the variance function:* (i) $v(\cdot) \in C_2([0,\infty))$, (ii) $\varsigma(\cdot)$ *is strictly increasing and strictly concave.*

Lemma 12.3.6 *Assumption 12.3.5 entails that, for any b, minimization* (12.1) *has a* unique *minimizer $t(b)$. In fact, $t(b)$ is the unique solution to*

$$F(b,t) := 2cv(t) - (b+ct)v'(t) = 0, \;\; or \;\; b = c\left(2\frac{v(t)}{v'(t)} - t\right). \tag{12.4}$$

Proof. First rewrite the minimization (12.1) as

$$\inf_{t\geq 0}\frac{m^2(t)}{2}, \;\; \text{with } m(t) := \frac{b+ct}{\varsigma(t)}.$$

Define $\phi(t) := \varsigma(t)/\varsigma'(t) - t$. Since

$$m'(t) = \frac{c\varsigma(t) - (b+ct)\varsigma'(t)}{\varsigma^2(t)},$$

and because of element (ii) of Assumption 12.3.5, it suffices to prove that (i) for each $b > 0$ and $c > 0$

$$\phi(t) = \frac{b}{c} \tag{12.5}$$

has a root $t(b)$, and (ii) $\phi(\cdot)$ is strictly increasing.

Due to $v(t)/t^\alpha \to 0$ for some $\alpha < 2$, it follows that $\lim_{t\to\infty} m(t) = \infty$ for each $b,c > 0$. Moreover, since $\varsigma(0) = 0$, it follows that $\lim_{t\to 0} m(t) = \infty$ for each $b,c > 0$. As a consequence, Equation (12.5) has at least one solution. Moreover

$$\phi'(t) = \frac{(\varsigma'(t))^2 - \varsigma(t)\varsigma''(t)}{(\varsigma'(t))^2} - 1 = -\frac{\varsigma(t)\varsigma''(t)}{(\varsigma'(t))^2} > 0,$$

since $\varsigma''(t) < 0$ due to the strict concavity of $\varsigma(\cdot)$, cf. element (ii) of Assumption 12.3.5. Thus $\phi(\cdot)$ is strictly increasing. This completes the proof. □

Our main result on the relation between the shape of the decay rate function $I_c^{(\mathrm{f})}(\cdot)$, and the correlation structure of the Gaussian sources, is stated in Theorem 12.3.9. We first prove two lemmas.

The first lemma says that the most likely epoch of overflow $t(b)$ is an increasing function of the buffer size b.

Lemma 12.3.7 $t(\cdot) \in C_1([0,\infty))$*, and is strictly increasing.*

Proof. Recall the fact that $t(b)$ is the *unique* solution to (12.4). In conjunction with $v(\cdot) \in C_2([0,\infty))$ and $v'(\cdot) > 0$ (Assumption 12.3.5), we conclude that $t(\cdot)$ is continuous. From (12.4), we see that

$$\begin{aligned} t'(b) = -\frac{\partial F/\partial b}{\partial F/\partial t} &= \frac{v'(t(b))}{cv'(t(b)) - (b+ct(b))v''(t(b))} \\ &= \frac{1}{c} \cdot \left(1 - 2\frac{v(t(b))v''(t(b))}{v'(t(b))^2}\right)^{-1}, \end{aligned} \tag{12.6}$$

such that the continuity of $t(\cdot)$, together with $v(\cdot) \in C_2([0,\infty))$, implies that $t'(\cdot)$ is continuous, too.

Assumption 12.3.5 states that, for all $t \geq 0$,

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\sqrt{v(t)} < 0 \iff 2\frac{v(t)v''(t)}{v'(t)^2} < 1,$$

thus proving the lemma. □

As we have seen in the proof of Thm. 12.3.3, $I_c^{(\mathrm{f})}(b)$ can be written as the variational problem

$$I_c^{(\mathrm{f})}(b) = \inf_{t\geq 0}\sup_{\theta} \left(\theta(b+ct) - \log \mathbb{E}e^{\theta A(t)}\right). \tag{12.7}$$

The optimizing θ reads

$$\theta_t(b) := \frac{b+ct}{v(t)}. \tag{12.8}$$

The second lemma states a relation between the derivative of the loss curve and the tilting parameter of the Fenchel-Legendre transform in (12.7). Here we use the shorthand notation $\theta(b) \equiv \theta_{t(b)}(b)$.

Lemma 12.3.8 *For all* $b > 0$*, it holds that* $(I_c^{(\mathrm{f})})'(b) = \theta(b)$.

Proof. Recalling that $t(b)$ is the optimizing t, differentiating (12.1) with respect to b yields

$$(I_c^{(\mathrm{f})})'(b) = \left(\frac{b+ct(b)}{v(t(b))}\right) - t'(b)\left(\frac{b+ct(b)}{2v^2(t(b))}\right)\left((b+ct(b))v'(t(b)) - 2cv(t(b))\right).$$

Now note that this equals $\theta(b)$, due to (12.4) and (12.8). □

The main result of this section can be proven now [13]. It describes the duality relation between the shape of $I_c^{(\mathrm{f})}(\cdot)$ and the correlation structure (which is uniquely determined by $v(\cdot)$). More specifically, it is shown that the curve $I_c^{(\mathrm{f})}(\cdot)$ is convex at some buffer size b if and only if there are negative correlations on the timescale $t(b)$ on which the overflow most likely takes place.

Theorem 12.3.9 *For all $b > 0$,*

$$(I_c^{(\mathrm{f})})''(b) \geq 0 \iff v''(t(b)) \leq 0.$$

Proof. Due to Lemma 12.3.8, $(I_c^{(\mathrm{f})})''(b) = \theta'(b)$. Trivial calculus yields

$$\theta'(b) = \frac{v(t(b))(1+ct'(b)) - 2ct'(b)v(t(b))}{v^2(t(b))} = \frac{1-ct'(b)}{v(t(b))},$$

where the last equality is due to (12.4). As $v(t)$ is nonnegative for any $t > 0$, conclude that $(I_c^{(\mathrm{f})})''(b) \geq 0$ is equivalent to $ct'(b) \leq 1$. So we are left to prove that $ct'(b) \leq 1$ is equivalent to $v''(t(b)) \leq 0$.

To show this equivalence, note that relation (12.6) yields

$$t'(b) = \frac{1}{c}\left(1 - \left(t(b) + \frac{b}{c}\right)\frac{v''(t(b))}{v'(t(b))}\right)^{-1}.$$

Now recall that $t(b) \geq 0$, $t'(b) \geq 0$ (due to Lemma 12.3.7) and $v'(t(b)) \geq 0$. Conclude that $ct'(b) \leq 1$ is equivalent to $v''(t(b)) \leq 0$. □

Obviously, for all b and t, it holds that $I_c^{(\mathrm{f})}(b) \leq (b+ct)^2/(2v(t))$. Noticing that both $v(\cdot)$ and $I_c^{(\mathrm{f})}(\cdot)$ are nonnegative, this results in the following interesting corollary.

Corollary 12.3.10 *For all $t > 0$, it holds that*

$$v(t) \leq \inf_{b>0} \frac{(b+ct)^2}{2I_c^{(\mathrm{f})}(b)}. \tag{12.9}$$

As described in [17], the inequality in (12.9) is, under mild conditions on $v(\cdot)$, actually an equality. This means that an interesting duality holds: when knowing the loss curve, one can retrieve the variance of the input process.

Example 12.3.11 – iOU. First verify that $v(t) = t - 1 + e^{-t}$ satisfies Assumption 12.3.5. It is easy to see that $v(\cdot)$ is convex, so we will have 'decreasing marginal buffering benefits', i.e., $I_c^{(\mathrm{f})}(\cdot)$ is concave due to Theorem 12.3.9. This example shows the relation between the 'level of positive correlation' and the shape of $I_c^{(\mathrm{f})}(\cdot)$. The strong convexity for small t indicates strong positive correlation on short timescale, whereas this positive correlation becomes weaker and weaker as the timescale increases (reflected by the asymptotically linear shape of $v(\cdot)$ for t large).

First we concentrate on small b. Straightforward calculations reveal that $I_c^{(\mathrm{f})}(b) = c^2 + \frac{2}{3}\sqrt{6bc^3} + O(b)$, where $t(b) \approx \sqrt{6b/c}$. So $I_c^{(\mathrm{f})}(\cdot)$ is highly concave for b small (i.e., behaving as a square root), expressing the strong positive correlations on a short timescale.

For large b, we find that $I_c^{(\mathrm{f})}(b) - 2c(b+c) \to 0$, with $t(b) \approx b/c + 2$. Apparently, for large b, $I_c^{(\mathrm{f})}(\cdot)$ becomes nearly linear, as expected by the weak correlation on long timescales. $\diamondsuit$

Example 12.3.12 – fBm. For $H < \frac{1}{2}$ this function is (uniformly) concave, indicating negative correlations, whereas $H > \frac{1}{2}$ entails that $v(\cdot)$ is convex corresponding to positive correlations — for $H = \frac{1}{2}$, the increments are independent. Assumption 12.3.5 is fulfilled; notice that $\sqrt{v(t)} = t^H$, which is concave. The results of Example 12.3.4 show that $I_c^{(\mathrm{f})}(\cdot)$ is indeed convex (concave) when the Hurst parameter is smaller (larger) than $\frac{1}{2}$, as could be expected on the basis of Theorem 12.3.9. $\diamondsuit$

12.3.4 The buffer-bandwidth curve is convex

A network provider has essentially two types of resources that he can deploy to meet the customers' performance requirements. When he chooses to increase the amount of buffer available in the network element, this clearly has a positive impact on the loss probability (albeit at the expense of incurring additional delay); the alternative is to increase the queue's service capacity (which reduces both the loss probability and the delay).

In other words: to achieve a certain predefined loss probability, say ε, the provider has to choose with which buffer size and link capacity this target is achieved. It is clear that the two types of resources trade off, and the goal of this section is to further analyze the corresponding buffer-bandwidth curve.

As before, we rely on the many-sources framework introduced earlier in this chapter: we have n sources sharing a network element with service rate nc and buffer threshold nb, with the performance objective $p_n(b,c) \leq \varepsilon$. Relying on the (very crude) approximation $p_n(b,c) \approx \exp(-nI_c^{(\mathrm{f})}(b))$, our objective becomes $I_c^{(\mathrm{f})}(b) \geq \delta$, where the identification $e^{-n\delta} = \varepsilon$ is used (such that $\delta > 0$). In other words: all values b,c such that

$$\inf_{t\geq 0} \frac{(b+ct)^2}{2v(t)} \geq \delta$$

satisfy the performance requirement.

Interestingly, the many-sources framework allows us to find the minimally required link capacity c for a given buffer b and loss constraint δ, as follows. By definition

$$c_b(\delta) \equiv c_b := \inf\left\{ c \mid \inf_{t\geq 0} \frac{(b+ct)^2}{2v(t)} \geq \delta \right\}.$$

It is clear, however, that if the infimum of a function $f(t)$ over t is larger than (or equal to) δ, then *for all t* it should hold that $f(t) \geq \delta$. In other words:

$$c_b = \inf\left\{c \mid \forall t \geq 0 : \frac{(b+ct)^2}{2v(t)} \geq \delta\right\}.$$

Isolating the c, this further reduces to

$$\begin{aligned} c_b &= \inf\left\{c \mid \forall t \geq 0 : c \geq \frac{\sqrt{2\delta v(t)} - b}{t}\right\} \\ &= \inf\left\{c \mid c \geq \sup_{t\geq 0} \frac{\sqrt{2\delta v(t)} - b}{t}\right\} = \sup_{t\geq 0} \frac{\sqrt{2\delta v(t)} - b}{t}. \end{aligned} \tag{12.10}$$

Similarly, the minimally required b (for given c, δ) can be computed:

$$b_c = \sup_{t\geq 0}\left(\sqrt{2\delta v(t)} - ct\right).$$

Interestingly, it now follows that the resources trade off in a convex way, in the sense that, for given δ, c_b is a convex function of b [11].

Proposition 12.3.13 *The required link capacity $c_b(\delta) \equiv c_b$ for given buffer b and decay rate δ, as given by* (12.10)*, is a convex function.*

Proof. Evidently, the objective function in (12.10), i.e., $\sqrt{2\delta v(t)}/t - b/t$, is linear in b. The maximum of linear functions is convex. □

Example 12.3.14 We now compute c_b and b_c for fBm. Applying the results above, we have that $c_b = \inf_{t\geq 0} f(t)$, with

$$f(t) := \sqrt{2\delta}\, t^{H-1} - \frac{b}{t}.$$

It is clear that $f(t) \to -\infty$ as $t \downarrow 0$; also $f(t) \to 0$ as $t \to \infty$. At the same time it can be verified that $f'(\cdot)$ has one zero, and $f''(\cdot)$ changes sign just once. In other words: we find the unique maximum by solving $f'(t) = 0$. This yields

$$t = \left(\frac{b}{\sqrt{2\delta}(1-H)}\right)^{1/H}.$$

Inserting this in the objective function yields

$$c_b = H(2\delta)^{1/2H}\left(\frac{b}{1-H}\right)^{1-1/H}.$$

We see that c and b trade off ‘hyperbolically’. Similarly,

$$b_c = (1-H)(2\delta)^{-1/(2H-2)}\left(\frac{c}{H}\right)^{H/(H-1)}.$$

The above calculations reveal that, along the trade-off curve, $b^{1-H}c^H$ remains constant (where this constant depends on H and δ). ◇

Example 12.3.15 We again consider fBm traffic, and require that the decay rate of the loss probability is at least δ. We impose the following cost structure: the cost per unit buffer is κ_b, and the cost per unit capacity is κ_c. The optimal buffer size $b^\star$ and capacity $c^\star$ are determined as follows.

We saw that, to obtain a decay rate δ, the resources b and c are such that $b^{1-H}c^H$ is constant; this constant, say φ, depends on H and δ. Consequently, the problem we have to solve is:

$$\min_{b\geq 0, c\geq 0} \kappa_b b + \kappa_c c \;\; \text{subject to} \;\; b^{1-H}c^H = \varphi.$$

Due to the convex form of the constraint, this can be solved immediately through Lagrangian optimization. It is easily verified that

$$c^\star = \left(\frac{\kappa_b}{\kappa_c}\cdot\frac{H}{1-H}\right)^{1-H}, \;\; b^\star = \left(\frac{\kappa_c}{\kappa_b}\cdot\frac{1-H}{H}\right)^{H}.$$

In fact, also for a general function $c_b(\delta)$ the solution can be characterized. Elementary convex analysis yields that we have to find the b for which the derivative of $c_b(\delta)$ is κ_b/κ_c, i.e., $b^\star$ is solved from

$$\frac{\kappa_b}{\kappa_c} = -\left(\frac{\partial}{\partial b}c_b(\delta)\right);$$

$c^\star$ then equals $c_{b^\star}(\delta)$. ◇

12.4 Tandem networks

Having focused on the single queue in the previous section, we now consider a two-queue tandem model, with (deterministic) service rate nc_1 for the first queue and nc_2 for the second queue. We assume that $c_1 > c_2$, in order to exclude the trivial case where the buffer of the second queue cannot build up.

12.4.1 Alternative formulation

In line with the previous section, we consider n i.i.d. Gaussian sources that feed into the first queue. Traffic of these sources that has been served at the first queue immediately flows into the second queue — we assume no additional sources to feed the second queue. We are interested in the steady-state probability of the buffer content of the second queue $Q_{2,n}$ exceeding a certain threshold nb, $b > 0$, when the number of sources gets large, or, more specifically, its logarithmic asymptotics:

$$I_c^{(\mathrm{t})}(b) := -\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(Q_{2,n} > nb), \tag{12.1}$$

where c denotes the vector $(c_1, c_2)^{\mathrm{T}}$. Note that we assume the buffer sizes of both queues to be infinite.

We first show that the probability of our interest can be written in terms of the 'empirical mean process' $n^{-1}\sum_{i=1}^n A_i(\cdot)$. The following lemma exploits the fact that we know both a representation of the first queue $Q_{1,n}$ (in steady-state) and a representation of the *total* queue $Q_{1,n} + Q_{2,n}$ (in steady-state). Let $t^0 := b/(c_1 - c_2)$.

Lemma 12.4.1 *$\mathbb{P}(Q_{2,n} > nb)$ equals*

$$\mathbb{P}\left(\exists t > t^0 : \forall s \in (0,t) : \frac{1}{n}\sum_{i=1}^n A_i(-t,-s) > b + c_2 t - c_1 s\right).$$

Proof. Notice that a 'reduction principle' applies: the total queue length is unchanged when the tandem network is replaced by its slowest queue, see e.g. [3, 8]. More formally: $Q_{1,n} + Q_{2,n} = \sup_{t>0}(\sum_{i=1}^n A_i(-t,0) - nc_2 t)$. Consequently we can rewrite the buffer content of the downstream queue as

$$\begin{aligned} Q_{2,n} &= (Q_{1,n} + Q_{2,n}) - Q_{1,n} \\ &=_{\mathrm{d}} \sup_{t>0}\left(\sum_{i=1}^n A_i(-t,0) - nc_2 t\right) - \sup_{s>0}\left(\sum_{i=1}^n A_i(-s,0) - nc_1 s\right). \end{aligned} \tag{12.2}$$

It was shown, see [23, Lemma 5.1], that the negative of the optimizing t in (12.2) corresponds to the start of the last busy period of the total queue in which time 0 is contained; similarly, the optimizing s is the start of the last busy period of the first queue in which time 0 is contained. Notice that a positive first queue induces a positive total queue, which immediately implies that we can restrict ourselves to $s \in (0,t)$. Hence $\mathbb{P}(Q_{2,n} > nb)$ equals

$$\mathbb{P}\left(\exists t > 0 : \forall s \in (0,t) : \frac{1}{n}\sum_{i=1}^n A_i(-t,-s) > b + c_2 t - c_1 s\right).$$

Because for $s \uparrow t$ the requirement

$$\frac{1}{n}\sum_{i=1}^n A_i(-t,-s) > b + c_2 t - c_1 s$$

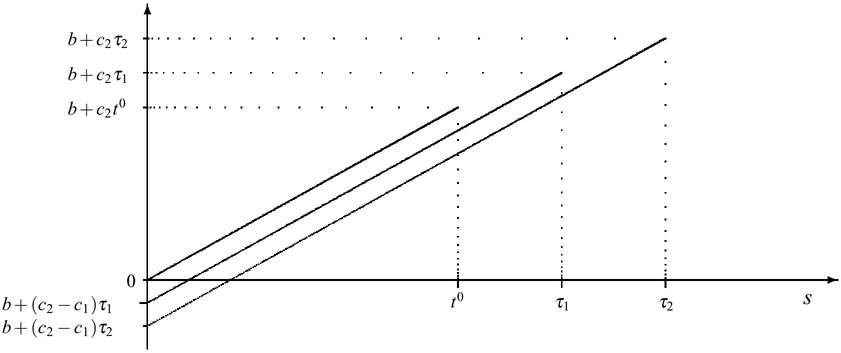


Fig. 12.1: Graphical representation of the overflow set. For different values of t, the curve $b+c_2t-c_1(t-s)$ has been drawn. Overflow occurs if there is a $t>t^0$ such that the empirical mean process lies, for $s\in(0,t)$, above the corresponding curve.

reads $0>b+(c_2-c_1)t$, we can restrict ourselves to $t>t^0$. We can interpret t^0 as the minimum time it takes to cause overflow in the second queue (notice that the maximum net input rate of the second queue in a tandem system is c_1-c_2). □

The crucial implication of the above lemma is that for analyzing $\mathbb{P}(Q_{2,n}\geq nb)$, we only have to focus on the behavior of the empirical mean process. More concretely,

$$\mathbb{P}(Q_{2,n}>nb)=p_n[\mathcal{S}^{(\mathrm{t})}]=\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}A_i(\cdot)\in\mathcal{S}^{(\mathrm{t})}\right), \tag{12.3}$$

where the set of 'overflow paths' $\mathcal{S}^{(\mathrm{t})}$ is given by

$$\mathcal{S}^{(\mathrm{t})}:=\{f\in\Omega:\exists t>t^0,\forall s\in(0,t):f(-s)-f(-t)>b+c_2t-c_1s\}.$$

Remark 12.4.2 A straightforward time-shift shows that the probability that the empirical mean process is in $\mathcal{S}^{(\mathrm{t})}$ coincides with the probability that it is in $\mathcal{T}$, with

$$\mathcal{T}:=\{f\in\Omega:\exists t>t^0,\forall s\in(0,t):f(s)>b+c_2t-c_1(t-s)\}. \tag{12.4}$$

However, the set $\mathcal{T}$ is somewhat easier to interpret, see Figure 12.1. For different values of t (i.e., $\tau_2>\tau_1>t^0=b/(c_1-c_2)$), the line $b+c_2t-c_1(t-s)$ has been drawn. The empirical mean process $n^{-1}\sum_{i=1}^{n}A_i(\cdot)$ is in $\mathcal{T}$ if there is a $t>t^0$ such that for all $s\in(0,t)$ it stays above the line $b+c_2t-c_1(t-s)$. Notice that $\mathcal{T}$ resembles the set corresponding to the probability of long busy periods in a single queue, as studied in [21]. ◇

Remark 12.4.3 As indicated above, our results are for centered sources, but, as before, they can be translated easily into results for non-centered sources, cf. Remark 12.3.2. Then the traffic generated by Gaussian source i in the interval $[s,t)$ is $A(s,t)+\mu(t-s)$, where $A(s,t)$ corresponds to a centered source; here $0<\mu<\min\{c_1,c_2\}$ and $s<t$. Let $q(\mu,c_1,c_2)$ be the probability that the second queue exceeds nb, given that input rate μ and service rates c_1 and c_2 are in force. From (12.2) it follows immediately that

$$q(\mu,c_1,c_2)=q(0,c_1-\mu,c_2-\mu),$$

and hence we can restrict ourselves to centered sources. ♢

12.4.2 Lower bound

In this section we start analyzing the logarithmic asymptotics of $\mathbb{P}(Q_{2,n}>nb)$. More specifically, we use 'Schilder' (Theorem 12.2.7) to formulate the decay rate as a variational problem, and then we find a lower bound on this decay rate.

Decay rate as a variational problem. We now consider the decay rate (12.1) of $\mathbb{P}(Q_{2,n}>nb)$. We already saw in Equation (12.3) that $\mathbb{P}(Q_{2,n}>nb)$ can be rewritten as the probability that the empirical mean process is in $\mathcal{S}^{(\mathrm{t})}$ (which is an open subset of Ω). The existence of the decay rate is now a consequence of Schilder's theorem, by showing (the plausible fact) that $\mathcal{S}^{(\mathrm{t})}$ is an $\mathbb{I}$-continuity set, i.e., that the infima of $\mathbb{I}(\cdot)$ over $\mathcal{S}^{(\mathrm{t})}$ and its closure, say $\overline{\mathcal{S}^{(\mathrm{t})}}$, match. This proof of $\mathcal{S}^{(\mathrm{t})}$ being an $\mathbb{I}$-continuity set is beyond the scope of this chapter, and can be found in Appendix A of [18].

Theorem 12.4.4

$$I_c^{(\mathrm{t})}(b)=\inf_{f\in\overline{\mathcal{S}^{(\mathrm{t})}}}\mathbb{I}(f)=\inf_{f\in\mathcal{S}^{(\mathrm{t})}}\mathbb{I}(f).$$

Lower bound on the decay rate. Our next goal is to derive a tractable lower bound on $I_c^{(\mathrm{t})}(b)$. This is presented in Theorem 12.4.5.

Observe that

$$\mathcal{S}^{(\mathrm{t})}=\bigcup_{t>t^0}\bigcap_{s\in(0,t)}\mathcal{S}_{s,t}^{(\mathrm{t})}\ \text{ with }\ \mathcal{S}_{s,t}^{(\mathrm{t})}:=\{f\in\Omega: f(-s)-f(-t)>b+c_2t-c_1s\}.$$

Hence we are interested in the decay rate of the union of intersections. The decay rate of a union of events is simply the minimum of the decay rates of the individual events, as we have seen several times before. The decay rate of an intersection, however, is not standard. In the next theorem we find a straightforward lower bound on this decay rate. Define

$$\mathcal{U}_{s,t}:=\{f\in\Omega: -f(-t)\geq b+c_2t; f(-s)-f(-t)\geq b+c_2t-c_1s\}.$$

Theorem 12.4.5 *The following lower bound applies:*

$$I_c^{(\mathrm{t})}(b) \geq \inf_{t>t^0} \sup_{s\in(0,t)} \inf_{f\in\mathcal{U}_{s,t}} \mathbb{I}(f). \tag{12.5}$$

Proof. Clearly,

$$I_c^{(\mathrm{t})}(b) = \inf_{t>t^0} \inf_{f\in \bigcap\limits_{s\in(0,t)} \mathcal{S}_{s,t}^{(\mathrm{t})}} \mathbb{I}(f).$$

Now fix t and consider the inner infimum. If $f(-s)-f(-t) > b+c_2t-c_1s$ for all $s\in(0,t)$, then also (f is continuous) $f(-s)-f(-t) \geq b+c_2t-c_1s$ for all $s\in[0,t]$. Hence,

$$\bigcap_{s\in(0,t)} \mathcal{S}_{s,t}^{(\mathrm{t})} \subseteq \bigcap_{s\in[0,t]} \mathcal{U}_{s,t} \subseteq \mathcal{U}_{r,t}$$

for all $r\in(0,t)$, and consequently

$$\inf_{f\in \bigcap\limits_{s\in(0,t)} \mathcal{S}_{s,t}^{(\mathrm{t})}} \mathbb{I}(f) \geq \inf_{f\in\mathcal{U}_{r,t}} \mathbb{I}(f).$$

Now take the supremum over r in the right-hand side. □

Theorem 12.4.5 contains an infimum over $f\in\mathcal{U}_{s,t}$. In the next lemma we show how this infimum can be computed.

Before stating this lemma, we first introduce some additional notation. Recalling 'bivariate Cramér' [6], the bivariate large-deviations rate function of

$$\left(\sum_{i=1}^{n} \frac{A_i(-t,0)}{n}, \sum_{i=1}^{n} \frac{A_i(-t,-s)}{n}\right)$$

is, for $y,z\in\mathbb{R}$ and $t>0$, $s\in(0,t)$, given by $\Lambda(y,z) := \frac{1}{2}(y,z)\,\Sigma(t-s,t)^{-1}(y,z)^{\mathrm{T}}$, with

$$\Sigma(s,t) := \begin{pmatrix} v(t) & \Gamma(s,t) \\ \Gamma(s,t) & v(s) \end{pmatrix}.$$

We also define the following quantity, which plays a key rôle in our analysis:

$$\begin{aligned} k(s,t) &:= \mathbb{E}(A(-s,0)\mid A(-t,0)=b+c_2t) \\ &= \mathbb{E}(A(s)\mid A(t)=b+c_2t) = \frac{\Gamma(s,t)}{v(t)}(b+c_2t). \end{aligned} \tag{12.6}$$

Recall Assumption 12.3.5: the standard deviation function was supposed to be $\mathcal{C}^2([0,\infty))$ and strictly increasing and strictly concave.

Lemma 12.4.6 *Under Assumption 12.3.5, for $t>t^0$ and $s\in(0,t)$,*

$$\inf_{f\in\mathcal{U}_{s,t}} \mathbb{I}(f) = \Upsilon(s,t) := \begin{cases} \Lambda(b+c_2t, b+c_2t-c_1s), & \text{if } k(s,t) > c_1s; \\ (b+c_2t)^2/2v(t), & \text{if } k(s,t) \leq c_1s. \end{cases}$$

Proof. Observe that

$$p_n[\mathfrak{U}_{s,t}] \equiv \mathbb{P}\left(\sum_{i=1}^{n}\frac{A_i(\cdot)}{n} \in \mathfrak{U}_{s,t}\right)$$
$$= \mathbb{P}\left(\sum_{i=1}^{n}\frac{A_i(-t,0)}{n} \geq b+c_2t; \sum_{i=1}^{n}\frac{A_i(-t,-s)}{n} \geq b+c_2t-c_1s\right). \quad (12.7)$$

We conclude that we can use 'bivariate Cramér' [6] to find the decay rate of $p_n[\mathfrak{U}_{s,t}]$. We obtain

$$\inf_{f\in\mathfrak{U}_{s,t}} \mathbb{I}(f) = \inf \Lambda(y,z),$$

where the last infimum is over $y \geq b+c_2t$ and $z \geq b+c_2t-c_1s$. Using that $\Lambda(\cdot,\cdot)$ is convex, this problem can be solved in the standard manner. It is easily verified that the contour of Λ that touches the line $y = b+c_2t$ does so at z-value

$$z^\star := \frac{\Gamma(t-s,t)}{v(t)}(b+c_2t);$$

also, the contour that touches $z = b+c_2t-c_1s$ does so at y-value

$$y^\star := \frac{\Gamma(t-s,t)}{v(t-s)}(b+c_2t-c_1s).$$

We first show that it cannot be that $y^\star > b+c_2t$, as follows. If $y^\star > b+c_2t$, then the optimum would be attained at $(y^\star, b+c_2t-c_1s)$. Straightforward computations, however, show that $y^\star > b+c_2t$ would imply that (use $\Gamma(t,t-s) \leq \sqrt{v(t)v(t-s)}$)

$$\left(\sqrt{v(t)} - \sqrt{v(t-s)}\right)(b+c_2t) > \sqrt{v(t)}\,c_1s. \quad (12.8)$$

This inequality is not fulfilled for $s = 0$ ($0 \not> 0$) nor for $s = t$ ($b+c_2t \not> c_1t$ for $t > t_0$). As the left hand side of (12.8) is convex (in s) due to Assumption 12.3.5, whereas the right hand is linear (in s), there is no $s \in (0,t)$ for which the inequality holds. Conclude that $y^\star > b+c_2t$ can be ruled out.

Two cases are left:

A) Suppose $z^\star > b+c_2t-c_1s$, or, equivalently, $k(s,t) \leq c_1s$. Then $(b+c_2t, z^\star)$ is optimal, with rate function $(b+c_2t)^2/2v(t)$, independent of s.

B) In the remaining case (where $y^\star \leq b+c_2t$ and $z^\star \leq b+c_2t-c_1s$) the optimum is attained at the $(b+c_2t, b+c_2t-c_1s)$, i.e., the 'corner point'. This happens if $k(s,t) > c_1s$, and gives the desired decay rate.

This proves the stated. As an aside we mention that if $k(s,t) = c_1s$, then both regimes coincide: $\Lambda(b+c_2t, b+c_2t-c_1s) = (b+c_2t)^2/2v(t)$. □

Summarizing, we have shown that, under Assumption 12.3.5, the following lower bound applies:

$$I_c^{(\mathrm{t})}(b) \geq \inf_{t>t^0} \sup_{s\in(0,t)} \Upsilon(s,t). \tag{12.9}$$

12.4.3 Tightness; two regimes

For large values of c_1 one would expect that the traffic characteristics are hardly changed by traversing the first queue. Define

$$L_c(t) := \frac{(b+ct)^2}{2v(t)},$$

and let t_c denote a t for which $L_c(t)$ is minimized. Then [18] shows that there is a critical link rate

$$c_1^\star := \sup_{s\in(0,t_{c_2})} \frac{k(s,t_{c_2})}{s},$$

above which the tandem system essentially behaves as a single queue, as formalized in the following result.

Theorem 12.4.7 *Under Assumption 12.3.5, if $c_1 \geq c_1^\star$, then*

$$I_c^{(\mathrm{t})}(b) = \inf_{t>t^0} \sup_{s\in(0,t)} \Upsilon(s,t) = L_{c_2}(t_{c_2}).$$

Remark 12.4.8 The approach we follow in this section to analyze the two-node tandem network, can be easily adapted to the setting of an m-node tandem network, with strictly decreasing service rates, i.e., $c_1 > \ldots > c_m$ — nodes i for which $c_i \leq c_{i+1}$ can be ignored, cf. [3]. Note that $\sum_{i=1}^k Q_{i,n}$ is equivalent to the single queue in which the sources feed into a buffer that is emptied at rate c_k. This means that we have the characteristics of both $\sum_{i=1}^{m-1} Q_{i,n}$ and $\sum_{i=1}^{m} Q_{i,n}$, which enables the analysis of $Q_{m,n}$, just as in the two-node tandem case. ♢

As shown in [15, 18], for iOU input the lower bound (12.9) is actually tight (and $c_1 \leq c_1^\star$), but this is not the case for fBm. The difficulty when looking for the most likely path is that, for fixed t, we have to deal with an infinite intersection of events, indexed by $s \in (0,t)$. We found a lower bound on the decay rate of the intersection, which corresponded to the least likely event in the intersection. As remarked earlier, the lower bound is tight if this least likely event essentially implies the other events in the intersection. Apparently, for iOU this is the case (and is the most likely path equivalent to the weighted sum of two covariance functions), but for fBm it is not.

The question remained what the most likely path should be for fBm. To investigate this issue, [16] studied first at a more elementary case. Consider the decay rate of $p_n[\mathcal{V}]$, with

$$\mathcal{V} := \bigcap_{t\in(0,1)} \mathcal{V}_t \text{ with } \mathcal{V}_t := \{f \in \Omega : f(t) \geq t\},$$

which could be interpreted as the event of having a busy period of length at least 1, in a queue drained at unit rate. Norros [21] already provided several bounds on this decay rate:

- as $\mathcal{V}_t \subseteq \mathcal{V}$, we have

$$-\lim_{n\to\infty} \frac{1}{n} \log p_n[\mathcal{V}] \geq \sup_{t\in(0,1)} \frac{t^2}{2t^{2H}} = \frac{1}{2};$$

- as the norm of any feasible path is an upper bound, we have

$$-\lim_{n\to\infty} \frac{1}{n} \log p_n[\mathcal{V}] \leq \mathbb{I}(\chi) =: \vartheta(H),$$

  where $\chi(t) = t$, for $t \in (0,1)$.

The function $\vartheta(\cdot)$ could be evaluated explicitly, and numerical investigations indicated that there was still a modest gap between the lower bound (i.e., $\frac{1}{2}$) and the upper bound. In [16] the exact value for the decay rate was found. Notably, the most likely path $f^\star$ is for $H \in (\frac{1}{2}, 1)$ such that $f^\star(t) = t$ for $t \in [0,\tau] \cup \{1\}$ (where $\tau < \frac{1}{2}$), and that $f^\star(t) > t$ for $t \in (\tau, 1)$; similarly, in the regime $H \in (0, \frac{1}{2})$, we have that $f^\star(t) = t$ for $t \in \{0\} \cup [\tau, 1]$. Interestingly, the most likely path is now a linear combination of uncountably many covariance functions. The analysis is substantially more involved than that of this chapter, but a number of concepts could be still used; more specifically, the concept of least likely events turned out to be very useful. The results of [16] also show that the 'smoothness' of the Gaussian process under consideration plays an important rôle here, which also explains why for iOU the lower bound (12.9) was tight, but for fBm not.

The above analysis for busy periods can be extended to tandem networks fed by fBm, as is done in [15]. For $c_1 < c_1^\star$, a part of the most likely path is linear (just as for the busy-period problem described above).

12.4.4 Approximation

Interestingly, along the lines of Mannersalo and Norros [19, 20] also the following approximation can be proposed:

$$I_c^{(\mathrm{t})}(b) \approx \inf_{t\geq t^0} \frac{(b+c_2t)^2}{2v(t)}; \tag{12.10}$$

in [19, 20] this is called a (*rough*) *full link approximation*. The idea behind this approximation is the following. If $t_{c_2} \geq t^0$, and traffic has been generated at a more or less constant rate in $[-t_{c_2},0]$, then no (or hardly) traffic is left in the first queue at time 0, and the approximation seems reasonably accurate. If on the other hand $t_{c_2} < t^0$, then there will be traffic left in the first queue at time 0, so the input rate needs to be pushed down; therefore $t = t^0$ has to be chosen, such that the sources are forced to transmit at about rate c_1, and the first queue remains (nearly) empty. Numerical experiments have indicated that this approximation is quite accurate, see [15].

Example 12.4.9 – fBm. Choosing $v(t) = t^{2H}$ gives

$$t_{c_2} = \frac{b}{c_2}\frac{H}{1-H}.$$

By Theorem 12.4.7,

$$I_c^{(\mathrm{t})}(b) = \frac{1}{2}\left(\frac{b}{1-H}\right)^{2-2H}\left(\frac{c_2}{H}\right)^{2H}$$

for all $c_1 \geq c_1^\star$. Unfortunately, for general H there does not exist a closed-form expression for $c_1^\star$. Straightforward calculus yields that

$$c_1^\star = \frac{c_2}{2H}\left(\sup_{\alpha\in(0,1)} \frac{1+\alpha^{2H}-(1-\alpha)^{2H}}{\alpha}\right);$$

observe that in this case $c_1^\star$ does *not* depend on b. It can be verified that $c_1^\star$ is close to (i.e., slightly larger than) c_2/H.

Now turn to the case $c_1 < c_1^\star$. It is readily verified that $t_c < t^0$ corresponds to $c_1 < c_2/H$. We obtain

$$I_c^{(\mathrm{t})}(b) \approx \frac{1}{2}\left(\frac{b}{c_1-c_2}\right)^{2-2H} c_1^2,$$

based on the rough fill link approximation. ◇

12.5 Priority queues

In the previous section we analyzed overflow in the second queue of a tandem system. This analysis was enabled by the fact that we had explicit knowledge of both the *first* queue and the *total* queue. In the present section we use the same type of arguments to solve the (two-queue) priority system.

12.5.1 Lower bound

We consider a priority system with a link of capacity nc, fed by traffic of two classes, each with its own queue. Traffic of class 1 does not 'see' class 2 at all, and consequently we know how the *high-priority* queue $Q_{h,n}$ behaves. Also, due to the work-conserving property of the system, the *total* queue length $Q_{h,n}+Q_{\ell,n}$ can be characterized. Now we are able, applying the same arguments as for the tandem queue, to analyze the decay rate of the probability of exceeding some buffer threshold in the low-priority queue. This similarity between tandem and priority systems has been observed before, see for instance [7].

We let the system be fed by n i.i.d. high-priority (hp) sources, and an equal number of i.i.d. low-priority (lp) sources; both classes are independent. We assume that both hp and lp sources are Gaussian. Define the means by μ_h and μ_ℓ, and the variance functions by $v_h(\cdot)$ and $v_\ell(\cdot)$, respectively; also $\mu := \mu_h+\mu_\ell$ (where $\mu < c$) and $v(\cdot) := v_h(\cdot)+v_\ell(\cdot)$. We note that in this priority setting we cannot restrict ourselves to centered processes. We denote the amount of traffic from the i-th hp source in $(s,t]$, with $s<t$, by $A_{h,i}(s,t)$; we define $A_{\ell,i}(s,t)$ analogously. Also $\Gamma_h(s,t),\Gamma_\ell(s,t)$ and R_h,R_ℓ are defined as before.

Remark 12.5.1 Notice that this setting also covers the case that the number of sources of both classes are *not* equal. Assume for instance that there are $n\alpha$ lp sources. Multiplying μ_ℓ and $v_\ell(\cdot)$ by α and applying the fact that the Normal distribution is infinitely divisible, we arrive at n i.i.d. sources. ◇

In the tandem situation we could, without loss of generality, center the Gaussian sources. It can be checked easily that such a reduction property does not hold in the priority setting, since there is no counterpart of Remark 12.4.3. Hence we cannot assume without loss of generality that $\mu_h=\mu_\ell=0$.

Analogously to Lemma 12.4.1, we obtain that $\mathbb{P}(Q_{\ell,n}>nb)$ equals

$$\mathbb{P}\left(\exists t>0:\forall s>0:\frac{1}{n}\sum_{i=1}^{n}A_{h,i}(-t,-s)+\frac{1}{n}\sum_{i=1}^{n}A_{\ell,i}(-t,0)>b+c(t-s)\right).$$

Let $I_c^{(\mathrm{p})}(b)$ be the exponential decay rate of $\mathbb{P}(Q_{\ell,n}>nb)$; analogously to Theorem 12.4.4 it can be shown that this decay rate exists. Similarly to the tandem case, with $f(\cdot)\equiv(f_h(\cdot),f_\ell(\cdot))$,

$$\mathcal{S}_{s,t}^{(\mathrm{p})} := \{f\in\Omega\times\Omega : f_h(-s)-f_h(-t)-f_\ell(-t)>b+c(t-s)\};$$

$$\mathcal{U}_{s,t}^{(\mathrm{p})} := \left\{f\in\Omega\times\Omega : \begin{array}{c}-f_h(-t)-f_\ell(-t)\geq b+ct;\\ f_h(-s)-f_h(-t)-f_\ell(-t)\geq b+c(t-s)\end{array}\right\}; \qquad (12.1)$$

$$\mathbb{P}(Q_{\ell,n}>nb)=\mathbb{P}\left(\left(\frac{1}{n}\sum_{i=1}^{n}A_{h,i}(\cdot);\frac{1}{n}\sum_{i=1}^{n}A_{\ell,i}(\cdot)\right)\in\bigcup_{t>0}\bigcap_{s>0}\mathcal{S}_{s,t}^{(\mathrm{p})}\right).$$

Theorem 12.5.2 *The following lower bound applies:*

$$I_c^{(\mathrm{p})}(b) \geq \inf_{t>0} \sup_{s>0} \inf_{f \in \mathcal{U}_{s,t}^{(\mathrm{p})}} \mathbb{I}(f), \tag{12.2}$$

with $\bar{f}_h(t) := f_h(t) - \mu_h t$, $\bar{f}_\ell(t) := f_\ell(t) - \mu_\ell t$, *and*

$$\mathbb{I}(f) := \frac{1}{2}||\bar{f}_h||_{R_h}^2 + \frac{1}{2}||\bar{f}_\ell||_{R_\ell}^2.$$

12.5.2 Tightness; two regimes

The infimum over $f \in \mathcal{U}_{s,t}^{(\mathrm{p})}$ can be computed explicitly, as in Lemma 12.4.6. As the analysis is analogous to the tandem case, but the expressions are more complicated, we only sketch the procedure. Again there is a regime in which one of the two constraints is redundant. Define

$$k_p(s,t) := \mathbb{E}(A_h(s) \mid A_h(t) + A_\ell(t) = b + ct).$$

Using the convexity of the large-deviations rate function, it can be shown that, if

$$\mathbb{E}(A_h(t-s) + A_\ell(t) \mid A_h(t) + A_\ell(t) = b + ct) > b + c(t-s),$$

only the first constraint in (12.1) is tightly met; it is equivalent to require that $k_p(s,t) < cs$. (If $k_p(s,t) \geq cs$ either both constraints in (12.1) are met with equality, or only the second constraint is met with equality; exact conditions for these two cases are easy to derive, but these are not relevant in this discussion.) As before, under $k_p(s,t) < cs$, we obtain the decay rate

$$\inf_{f \in \mathcal{U}_{s,t}^{(\mathrm{p})}} \mathbb{I}(f) = \frac{(b + (c-\mu)t)^2}{2v(t)}, \tag{12.3}$$

cf. the single queue with link rate nc; in the other cases the expressions are somewhat more involved. Denote (in this section) by t_c the value of $t > 0$ that minimizes the right hand side of (12.3).

Similarly to the tandem case, there is a regime (i.e., a set of values of the link rate c) in which $I_c^{(\mathrm{p})}(b)$ coincides with the decay rate of a single queue. In this regime, which we call regime (A), conditional on a large value of the total queue length, it is likely that the hp queue is empty, such that all traffic that is still in the system is in the lp queue. Hence, for all c in

$$\{c \mid \forall s > 0 : k_p(s,t_c) < cs\} \tag{12.4}$$

we conclude

$$I_c^{(\mathrm{p})}(b) = \frac{(b+(c-\mu)t_c)^2}{2v(t_c)}.$$

If c is not in the set (12.4), a condition can be found [18] under which the lower bound of Theorem 12.5.2 is tight; we call this regime (B).

In the tandem case, we found that the single-queue result holds for $c_1 \geq c_1^\star$, whereas it does not hold for $c_1 < c_1^\star$; the threshold value $c_1^\star$ was found explicitly in Section 12.4.2. In the priority setting there is not such a clear dichotomy. Consider for instance the situation in which both types of sources correspond to Brownian motions; $v_h(t) \equiv \lambda_h t$, $v_\ell(t) \equiv \lambda_\ell t$, and $\lambda := \lambda_h + \lambda_\ell$. Define

$$\Xi := \sqrt{\mu_\ell^2 + \frac{\lambda_\ell}{\lambda_h}(c-\mu_h)^2}.$$

Then straightforward calculus yields that for $(\lambda_h - \lambda_\ell)c \leq \lambda_h(\mu_h + 2\mu_\ell) - \lambda_\ell\mu_h$, regime (A) applies (i.e., the single-queue result holds):

$$I_c^{(\mathrm{p})}(b) = \frac{2b(c-\mu)}{\lambda},$$

whereas otherwise we are in regime (B):

$$I_c^{(\mathrm{p})}(b) = \frac{b(\Xi-\mu_\ell)}{\lambda_\ell};$$

this is shown by verifying that the lower bound of Theorem 12.5.2 is tight for the specific case of Brownian motion input. Using $\mu_h + \mu_\ell < c$, it can be verified easily that this implies that for $\lambda_h \leq \lambda_\ell$ the single-queue solution applies, whereas for $\lambda_h > \lambda_\ell$ only for

$$c \leq \frac{\lambda_h(\mu_h + 2\mu_\ell) - \lambda_\ell\mu_h}{\lambda_h - \lambda_\ell},$$

the single-queue solution applies.

12.5.3 Approximation

Our lower bound reads

$$I_{c,\mathrm{I}}^{(\mathrm{p})}(b) := \inf_{t>0}\sup_{s>0} \Upsilon_p(s,t), \quad \text{with } \Upsilon_p(s,t) := \inf_{f\in\mathcal{U}_{s,t}^{(\mathrm{p})}} \mathbb{I}(f).$$

Just as we did above, Mannersalo and Norros [19] identify two cases. The same solution is obtained for our regime (A), i.e., the situation in which, given a long total queue length, the hp queue is relatively short. In regime (B) the hp queue tends to be large, given that the total queue is long. To prevent this from happening,

[19] proposes a heuristic (in line with the rough full link approximation that we introduced for the tandem case) that minimizes $\mathbb{I}(f)$ over

$$\{f \in \Omega \times \Omega : \exists t > 0 : -f_h(-t) - f_\ell(-t) \geq b + ct; -f_h(-t) \leq ct\}. \tag{12.5}$$

Because regime (B) applies, the optimum paths in the set (12.5) are such that the constraints on f are tightly met; consequently (12.5) is a subset of $\mathfrak{U}_{t,t}^{(\mathrm{p})}$. Hence the resulting decay rate, which we denote by $I_{c,\mathrm{II}}^{(\mathrm{p})}(b)$, yields a lower bound, but our lower bound will be closer to the real decay rate:

$$I_{c,\mathrm{II}}^{(\mathrm{p})}(b) := \inf_{t>0} \Upsilon_p(t,t) \leq \inf_{t>0} \sup_{s>0} \Upsilon_p(s,t) = I_{c,\mathrm{I}}^{(\mathrm{p})}(b).$$

Remark 12.5.3 In the simulation experiments performed in [19], it was found that the lower bound $I_{c,\mathrm{II}}^{(\mathrm{p})}(b)$ is usually close to the exact value. Our numerical experiments show that the hp buffer usually starts to fill shortly after the total queue starts its busy period. This means that in many cases the error made by taking $s = t$ is relatively small. It explains why the heuristic based on set (12.5) performs well. $\diamondsuit$

12.6 Concluding remarks

In this chapter we have described a family of results on queueing networks with Gaussian inputs. Under the many-sources scaling we have characterized the decay rate of exceeding a predefined threshold. The type of networks is still rather limited; future work may focus on a broad classes of networks (tree networks, tandem networks with exogenous input in the downstream queue, etc.); the structural results of [26, 27] could be useful here.

References

1. R. Addie, P. Mannersalo, and I. Norros. Most probable paths and performance formulae for buffers with Gaussian input traffic. *European Trans. Telecommun.*, 13:183–196, 2002.
2. R. Adler. *An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes. Lecture Notes-Monograph Series, Vol. 12.* Institute of Mathematical Statistics, Hayward, CA, USA, 1990.
3. B. Avi-Itzhak. A sequence of service stations with arbitrary input and regular service times. *Man. Sci.*, 11:565–571, 1965.
4. D. Botvich and N. Duffield. Large deviations, the shape of the loss curve, and econimies of large scale multiplexers. *Queueing Syst.*, 20:293–320, 1995.
5. C. Courcoubetis and R. Weber. Buffer overflow asymptotics for a buffer handling many traffic sources. *J. Appl. Probab.*, 33:886–903, 1996.
6. A. Dembo and O. Zeitouni. *Large deviations techniques and applications, 2nd edition*. Springer, New York, NY, USA, 1998.

7. A. Elwalid and D. Mitra. Analysis, approximations and admission control of a multi-service multiplexing system with priorities. In *Proc. IEEE Infocom*, pages 463–472, 1995.
8. H. Friedman. Reduction methods for tandem queueing systems. *Oper. Res.*, 13:121–131, 1965.
9. A. Ganesh, N. O'Connell, and D. Wischik. *Big Queues. Lecture Notes in Mathematics, Vol. 1838*. Springer, Berlin, Germany, 2004.
10. J. Kilpi and I. Norros. Testing the Gaussian approximation of aggregate traffic. In *Proc. 2nd Internet Measurement Workshop*, pages 49–61, 2002.
11. K. Kumaran and M. Mandjes. The buffer-bandwidth trade-off curve is convex. *Queueing Syst.*, 38:471–483, 2001.
12. W. Leland, M. Taqqu, W. Willinger, and D. Wilson. On the self-similar nature of Ethernet traffic (extended version). *IEEE/ACM Trans. Netw.*, 2:1–15, 1994.
13. M. Mandjes. A note on the benefits of buffering. *Stoch. Models*, 20:43–54, 2004.
14. M. Mandjes. *Large Deviations for Gaussian Queues*. Wiley, Chichester, UK, 2007.
15. M. Mandjes, P. Mannersalo, and I. Norros. Gaussian tandem queues with an application to dimensioning of switch fabrics. *Comp. Netw.*, 51:781–797, 2006.
16. M. Mandjes, P. Mannersalo, I. Norros, and M. van Uitert. Large deviations of infinite intersections of events in Gaussian processes. *Stoch. Proc. Appl.*, 116:1269–1293, 2006.
17. M. Mandjes and R. van de Meent. Resource provisioning through buffer sampling. *IEEE/ACM Trans. Netw.*, 2009.
18. M. Mandjes and M. van Uitert. Sample-path large deviations for tandem and priority queues with Gaussian inputs. *Ann. Appl. Probab.*, 15:1193–1226, 2005.
19. P. Mannersalo and I. Norros. Approximate formulae for Gaussian priority queues. In *Proc. ITC 17*, pages 991–1002, 2001.
20. P. Mannersalo and I. Norros. A most probable path approach to queueing systems with general Gaussian input. *Comp. Netw.*, 40:399–412, 2002.
21. I. Norros. Busy periods of fractional Brownian storage: a large deviations approach. *Adv. Perf. Anal.*, 2:1–20, 1999.
22. V. Paxson and S. Floyd. Wide area traffic: the failure of Poisson modeling. *IEEE/ACM Trans. Netw.*, 3:226–244, 1995.
23. E. Reich. On the integrodifferential equation of Takács I. *Ann. Math. Stat.*, 29:563–570, 1958.
24. R. van de Meent, M. Mandjes, and A. Pras. Gaussian traffic everywhere? In *Proc. IEEE International Conference on Communications*, pages 573–578, 2006.
25. A. Weiss. A new technique for analyzing large traffic systems. *Adv. Appl. Probab.*, 18:506–532, 1986.
26. D. Wischik. The output of a switch, or, effective bandwidths for networks. *Queueing Syst.*, 32:383–396, 1999.
27. D. Wischik and A. Ganesh. The calculus of Hurstiness. *Queueing Syst.* to appear

Chapter 13
Mean Values Techniques

Ivo Adan, Jan van der Wal

13.1 Introduction

This chapter presents the technique to determine mean performance characteristics of queueing models, generally known as *mean value analysis* (MVA). The term MVA is usually associated with queueing networks (QNs). However, the MVA technique is also very powerful when studying single-station queueing models. The merits of MVA are in its intrinsic simplicity and its intuitively appealing derivation based on probabilistic arguments. The intuition of MVA can also be used to develop approximations for problems where an exact analysis appears to be intractable.

We would like to emphasize that this chapter is not intended as an exhaustive survey of MVA. The main goal is to demonstrate the elegance and power of MVA for a collection of queueing problems and, so as to speak, to whet the reader's appetite to apply MVA to new problems.

MVA ideas for single stations must have been around for a long time, although it is hard to locate the exact origin. One of the reasons might be that the analysis of single-server stations is usually based on transform techniques, yielding mean values as a byproduct. MVA for QNs originated in the late seventies. The first MVA ideas were independently invented by Schweitzer [11] and Bard [2] and were approximate in nature. Shortly after the first approximate MVA techniques (AMVA) were developed, Lavenberg and Reiser [8, 9] discovered exact MVA. The advantages of exact and approximate MVA were recognized soon thereafter and the MVA literature for QNs grew rapidly.

The principle of MVA for single stations with Poisson arrivals is essentially based on two key properties: Little's law and the property that an arriving Poisson job finds the system in equilibrium. The latter property is commonly referred to as the PASTA

Ivo Adan, Jan van der Wal
Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands, and Department of Quantitative Economics, University of Amsterdam, Amsterdam, The Netherlands.
e-mail: {i.j.b.f.adan,jan.v.d.wal}@tue.nl

R.J. Boucherie and N.M. van Dijk (eds.), *Queueing Networks: A Fundamental Approach*,
International Series in Operations Research & Management Science 154,
DOI 10.1007/978-1-4419-6472-4_13, © Springer Science+Business Media, LLC 2011

property: Poisson Arrivals See Time Averages. In product-form (PF) networks the situation is similar. In open PF networks one again has these properties, although the PASTA property formally does not hold, as there need not be Poisson processes inside the network, but it is replaced by the equally powerful ASTA (Arrivals See Time Averages) property or so-called Arrival Theorem. In closed PF networks the arrival theorem is different: a job moving from one station to another does not find the system in equilibrium, but finds it in the equilibrium as if this job has never been in the system.

In the following sections we prefer to start with MVA ideas for single station systems. We will illustrate these ideas with a number of different applications demonstrating the ease and power of this way of thinking. It should be pointed out that a disadvantage of MVA for single stations (and in general) is that it provides *mean values only*, so no second or higher moments, let be distributions.

13.2 PASTA property and Little's law

The first key property that allows for a mean value approach is the fact that there is a close relation between the distribution of the state of the system at an arrival moment and the time average distribution of the system state. For Poisson arrivals this relation is the so-called PASTA property which is the acronym for: Poisson Arrivals See the state of the system as the Time Average. It can be intuitively explained by the memoryless property of the exponential inter-arrival times, as will be demonstrated below (see, e.g., Kleinrock [15], pp. 118-119; for a rigorous proof see Wolff [14]).

Let $X(t)$ denote the state of the system at time t; for example, $X(t)$ may indicate the number of jobs in the queue at time t. Further, let A be a subset of states. Then the probability that an arriving Poisson job in $(t, t+\Delta)$ finds the system in subset A is equal to

$$\begin{aligned}\Pr[X(t) \in A | \text{Arrival in}(t, t+\Delta t)] &= \frac{\Pr[X(t) \in A, \text{Arrival in}(t, t+\Delta t)]}{\Pr[\text{Arrival in}(t, t+\Delta t)]} \\ &= \frac{\Pr[X(t) \in A] \Pr[\text{Arrival in}(t, t+\Delta t) | X(t) \in A]}{\Pr[\text{Arrival in}(t, t+\Delta t)]}.\end{aligned}$$

Since Poisson arrivals are memoryless, we have

$$\Pr[\text{Arrival in}(t, t+\Delta t) | X(t) \in A] = \Pr[\text{Arrival in}(t, t+\Delta t)] ,$$

so that

$$\Pr[X(t) \in A | \text{Arrival in}(t, t+\Delta t)] = \Pr[X(t) \in A].$$

In the sequel this property will be used when the system is in equilibrium, thus as $t \longrightarrow \infty$. There is considerable freedom in the definition of *the state of the system*. For example, if it is defined as the number of jobs in the queue, then this property, applied to $A = \{0, 1, \ldots, n\}$, yields that the queue length distribution on arrival is equal to the equilibrium queue length distribution. Alternatively, if the state indicates the status of the server (1 whenever busy and 0 otherwise), then by applying this property to $A = \{1\}$, we obtain that the probability the server is busy on arrival is equal to the long-run fraction of time the server is busy.

The other key property is Little's law [16, 13] stating that

$$L = \lambda S,$$

where L is the mean number of jobs in the system, λ the arrival rate and S the mean sojourn time. Again, one can exploit the freedom in the definition of *the system* to obtain various relations. For example, in a station with a single queue and a single server, the system can be defined as the queue (so without the server), yielding the following relation between the mean queue length Q and the mean waiting time W,

$$Q = \lambda W. \tag{13.1}$$

But, when the system is defined as the server, we obtain a relation between the utilization ρ (i.e., fraction of time the server is busy) and the mean service time b,

$$\rho = \lambda b. \tag{13.2}$$

In the sequel, when writing down expressions on an arrival instant, we could write, for example, $\rho^{(a)}$ or $Q^{(a)}$ with a indicating the arrival, but because of the PASTA property these quantities are equal to the equilibrium quantities. Therefore we will not do this.

13.3 MVA for single-station systems

Below we present a couple of examples of single-station systems for which the MVA approach works; that is, for which performance characteristics, though their mean values only, can be obtained using PASTA and Little's law.

13.3.1 M|M|1

The simplest example of this principle is seen in the MVA approach for the $M|M|1$ queue with arrival rate λ and service rate μ, and in which jobs are served FCFS. For stability we assume that $\lambda < \mu$. Let Q denote the mean number of jobs in the queue and ρ the probability that the server is busy (*in equilibrium as well as on arrival*). The mean waiting time is W. Then we get

$$W = \rho \frac{1}{\mu} + Q \frac{1}{\mu}. \tag{13.1}$$

Note that, by the memoryless property of exponentials, the mean residual service time of the job in service upon arrival is also $1/\mu$. Relation (13.1) is usually referred to as the *arrival relation*. Together with Little's law (13.1), this results in

$$W = \frac{\rho}{1-\rho}\frac{1}{\mu},$$

where $\rho = \lambda/\mu$ by virtue of (13.2).

13.3.2 M|G|1

This approach can be easily extended to jobs requiring general service times with mean b and second moment $b^{(2)}$. For stability we require $\lambda b < 1$. Then, for the mean waiting time, we get

$$W = \rho R + Qb, \tag{13.2}$$

where R is the mean residual service time on arrival. Now we cannot conclude that R is equal to the mean service time b, as in the exponential case. According to the PASTA property R is equal to the (time) average residual service time R given by (see, e.g., Ross [17], pp. 424-425)

$$R = \frac{b^{(2)}}{2b} = \frac{b}{2}(1+c^2), \tag{13.3}$$

where $c^2 = (b^{(2)} - b^2)/b^2$ denotes the squared coefficient of variation of the service time. Together with Little's law (13.1) this immediately yields the Pollackzek-Khinchin formula,

$$W = \frac{\rho R}{1-\rho}. \tag{13.4}$$

Note that W can also be interpreted as the mean amount of work in the system.

13.3.3 M|G|1 *with priorities*

Let us now consider a single server treating C classes of jobs, labeled $1,\dots,C$. Class i jobs arrive according to a Poisson process at rate λ_i and require service times with mean b_i and second moment $b_i^{(2)}$. The class i utilization is $\rho_i = \lambda_i b_i$. Jobs are served according to *non-preemptive* priorities, i.e., class i has priority over all classes j with $j > i$, but service interruptions are not allowed. Per class the service discipline is assumed to be FCFS. For stability we require $\rho_1 + \cdots + \rho_C < 1$.

Let W_i denote the mean waiting time of class i jobs and Q_i the mean number of class i jobs waiting in the queue. Then the arrival relation for class i is

$$W_i = \sum_{j \le i} Q_j b_j + \sum_j \rho_j R_j + W_i \sum_{j<i} \lambda_j b_j \ , \tag{13.5}$$

where R_j is the mean residual service time of class j, so $R_j = b_j^{(2)}/2b_j$.

On the right-hand side of (13.5), the first term is the mean waiting time due to the higher or same priority jobs in the queue upon arrival. The second term is the mean amount of residual work in service and the third term captures the higher priority jobs that are expected to arrive and overtake the class-i job during its waiting time.

Using Little's law,

$$Q_j = \lambda_j W_j$$

and $\rho_j = \lambda_j b_j$, equation (13.5) can be rewritten as

$$W_i = \sum_{j \le i} \rho_j W_j + \sum_j \rho_j R_j + W_i \sum_{j<i} \rho_j \ ,$$

or

$$W_i(1 - \sum_{j \le i} \rho_j) = \sum_{j<i} \rho_j W_j + \sum_j \rho_j R_j. \tag{13.6}$$

For class 1 (for which there is no overtaking) this immediately results in

$$W_1 = \frac{\sum_j \rho_j R_j}{1 - \rho_1}. \tag{13.7}$$

Note that for $i > 1$ the right hand side of (13.6) can be written as

$$\rho_{i-1} W_{i-1} + \sum_{j<i-1} \rho_j W_j + \sum_j \rho_j R_j = \rho_{i-1} W_{i-1} + W_{i-1}(1 - \sum_{j \le i-1} \rho_j), \tag{13.8}$$

where, for the second and third term on the left, equation (13.6) is used with i replaced by $i - 1$. This leads to the recursion

$$W_i(1 - \sum_{j \le i} \rho_i) = W_{i-1}(1 - \sum_{j<i-1} \rho_j), \qquad i = 2, \ldots, C,$$

from which one easily gets

$$W_i = \frac{\sum_j \rho_j R_j}{(1 - \sum_{j \le i} \rho_j)(1 - \sum_{j<i} \rho_j)}, \qquad i = 1, 2, \ldots, C. \tag{13.9}$$

Note that (13.9) is indeed valid for $i = 1$, by virtue of (13.7).

13.3.4 M|G|1 *with least attained service*

In this section we consider a *dynamic* priority rule, the so-called Least Attained Service (LAS) policy: the server gives priority to the job that has received the least amount of service. Now we obviously allow service interruptions; e.g., an arriving job will immediately enter service, interrupting the one currently in service and whose service will resume at the point where it was interrupted as soon as it is the one again that has received the least amount of service. If the service time has a decreasing hazard rate, the LAS policy is known to minimize the mean waiting time among all policies which do not use information about the job sizes (see, e.g., Yashkov [19]).

For a job of size x, only jobs that have an attained service less or equal to x are visible. Hence, it experiences a single server treating jobs according to LAS with a *truncated* service time distribution

$$F_x(y) = \begin{cases} F(y) & \text{if } y < x, \\ 1 & \text{if } y \geq x, \end{cases}$$

where $F(\cdot)$ is the (original) service time distribution. Let b_x and $b_x^{(2)}$ denote the first two moments of the truncated service time distribution,

$$b_x = \int_0^x (1 - F(y))dy, \qquad b_x^{(2)} = \int_0^x 2y(1 - F(y))dy,$$

and $\rho_x = \lambda b_x$. The amount of work in this system does not depend on the order in which jobs are served, and thus it is the same as in the FCFS version. Thus, by (13.4), the mean amount of work is equal to

$$W_x = \frac{\rho_x R_x}{1 - \rho_x},$$

where $R_x = b_x^{(2)}/2b_x$. Note that, by the PASTA property, W_x is also the mean amount of work which an arriving job of size x finds in the system and which has to be served before that job leaves the system. Hence, for the mean sojourn time S_x of a job of size x we obtain, similar to (13.5),

$$S_x = x + W_x + \lambda S_x b_x,$$

where the last term is the mean amount of work that arrives during the sojourn time of the job of size x and which has to be served before that job leaves the system. Thus,

$$S_x = \frac{x + W_x}{1 - \rho_x},$$

and by unconditioning on the job size x, we get that the mean sojourn time S of an *arbitrary* job satisfies

$$S = \int_0^\infty S_x dF(x).$$

13.3.5 Server vacations

There are many models for server vacations. A simple one is the following. When the queue is empty after a service completion, the server takes a vacation. If upon return of the server the queue is still empty, the server immediately takes another vacation and otherwise the server starts servicing the queue.

Upon arrival the server is on vacation (i.e., not at work) with probability $1-\rho$, in which case the job has to wait for a residual vacation time. Let R_v denote the mean residual vacation time. Then the arrival relation is

$$W = (1-\rho)R_v + Qb + \rho R.$$

So, with Little's law, $Q = \lambda W$, one gets

$$W = \frac{\rho R}{1-\rho} + R_v. \tag{13.10}$$

Another simple vacation model is the following. When the server finishes the service of the last customer in a busy period he takes a vacation until the K-th new customer arrives. For example, the server could be an oven that is switched off partly if there is no work left. Note that in this case the duration of the vacation depends on the arrival times.

Clearly $1-\rho$ is the probability that a customer arrives during a server vacation. If so, the customer is with equal probability the first, second or K-th customer in that vacation period, so the residual duration of the vacation expressed in inter-arrival times, is with equal probability $K-1$, $K-2$, or 0, so an average $(K-1)/2$. This results in

$$W = (1-\rho)\frac{K-1}{2}\frac{1}{\lambda} + Qb + \rho R.$$

With Little's law one obtains

$$W = \frac{\rho R}{1-\rho} + \frac{K-1}{2}\frac{1}{\lambda}. \tag{13.11}$$

Note that the first term in both (13.10) and (13.11) is equal to the mean waiting time in the "ordinary" $M|G|1$ queue; see (13.4).

13.3.6 M|M|c

So far we considered single-server stations. Let us now consider an exponential multi-server queue, the $M|M|c$ with arrival rate λ and service rate μ. For stability we assume $\lambda < c\mu$.

If not all servers are busy on arrival, the waiting time is zero. If all servers are busy, the job has to wait until the first departure and then continues to wait for as many departures as there were jobs waiting upon arrival. An inter-departure time is the minimum of c exponentials with rate μ, and thus it is exponential with rate $c\mu$. The probability that all servers are busy (on arrival) is denoted by B (for busy say) and it is easily computed as

$$B = \frac{(c\rho)^c}{c!}\left((1-\rho)\sum_{n=0}^{c-1}\frac{(c\rho)^n}{n!} + \frac{(c\rho)^c}{c!}\right)^{-1}, \tag{13.12}$$

where $\rho = \lambda/c\mu$ denotes the server utilization. Hence, for the mean waiting time W, we obtain

$$W = B\frac{1}{c\mu} + Q\frac{1}{c\mu}. \tag{13.13}$$

Together with Little's law (13.1) this leads to

$$W = \frac{B}{1-\rho}\,\frac{1}{c\mu}.$$

13.3.7 **M|M|c** ***with priorities***

An extension of the previous model is the $M|M|c$ priority system, say without preemption, treating C classes of jobs. The classes are labeled $1,\ldots,C$. All jobs are statistically equal, i.e., all of them are exponential with the same mean $1/\mu$.

The waiting time of a class i job is determined by the number of jobs of the various classes found waiting upon arrival and also depends on whether upon arrival all servers are busy or not. The latter probability does not depend on the order in which the jobs are served, so it is equal to the probability B in case of the FCFS service discipline; see (13.12). This results in

$$W_i = B\frac{1}{c\mu} + \sum_{j\le i} Q_j\frac{1}{c\mu} + \sum_{j<i}\lambda_j W_i\frac{1}{c\mu}.$$

This expression is the same as (13.5) for the $M|G|1$ priority model, with b_j replaced by $1/(c\mu)$ and the term $\sum_j \rho_j R_j$ replaced by $B/c\mu$. Thus also result (13.9) is the same:

$$W_i = \frac{B}{(1-\sum_{j\le i}\rho_j)(1-\sum_{j<i}\rho_j)}\,\frac{1}{c\mu}, \qquad i = 1,\ldots,C.$$

In the models discussed sofar the MVA approach was rather straightforward, as only the arrival relation and Little's formula were used. In the next two sections we consider models for which a somewhat different reasoning is needed.

13.3.8 Retrials

There are many models where jobs finding a busy server on arrival leave immediately and retry later, see Artalejo [1]; think of, e.g., call centers. A basic *retrial* model is the following. Jobs arrive according to a Poisson process with rate λ at a single server. If the server is busy on arrival, the job immediately leaves and goes "in orbit" to return after an exponential time with mean $1/\gamma$, and if the server is still or again busy it goes back into orbit and so on.

The total time that a job is spending in orbit can be divided into two parts: time during which the server is working and time during which the server is idle and waiting for a new arrival, either one in orbit or one from the outside. During the orbit time the server completes on average $1+D$ services, where the first "1" is the job found in service upon arrival, which requires a residual service time R, and D is the (unknown) number of jobs that enter the server while the job under consideration is in orbit. During the server's idle time the job competes for the server; there are also attempts to find the server idle while the server is busy, but these can be ignored. So the retrial time is only counted when the server is idle and thus, by the memoryless property of exponentials, the retrial time during idle time is again exponential with mean $1/\gamma$. Hence, the mean sojourn time S (retrial time plus service time) satisfies

$$S = (1-\rho)\cdot 0 + \rho\cdot(R + Db + 1/\gamma) + b. \tag{13.14}$$

Further, in any queueing system with jobs arriving one by one and are leaving one by one, the number of jobs in the system just before an arrival and just after a departure is equally distributed. Hence, in equilibrium, the mean number of arrivals during the sojourn time is equal to the mean number of service completions during the sojourn time, so

$$\lambda S = (1-\rho)\cdot 0 + \rho\cdot(1+D). \tag{13.15}$$

Solving the equations (13.14) and (13.15) yields

$$S = \frac{\rho R}{1-\rho} + b + \frac{\rho}{1-\rho}\frac{1}{\gamma}$$

Again note that the first two terms are the mean sojourn time in the $M|G|1$ queue.

13.3.9 Polling

Another more complex model is the following *polling* model with a single server for N queues, labeled $1,\ldots,N$. Class i jobs arrive in queue i according to a Poisson process at rate λ_i and require service times with mean b_i and mean residual R_i. The queues are served exhaustively in the order $1,2,\ldots,N,1,\ldots$ and so on. When switching to queue i there is a switch-over time with mean s_i and mean residual R_{s_i}; the total switch-over time $s = \sum s_j$ is assumed to be positive. The fraction of time the server is busy with queue i is $\rho_i = \lambda_i b_i$, where, for stability, it is assumed that the total utilization $\rho = \sum_i \rho_i < 1$. The cycle length of queue i is the time between successive arrivals of the server to this queue. Its mean C is independent of i and satisfies $C = \rho C + s$, whence $C = s/(1-\rho)$. The visit time to queue i is defined as the switch-over time to queue i plus the time spent at queue i. Hence, the mean visit time V_i is equal to $s_i + \rho_i C$ for $i = 1, \ldots, N$. We will present the MVA approach for the case of two queues. The generalization to N queues is straightforward, but the notations become involved. Let us consider a job of class 1. Upon its arrival the job has to wait for all Q_1 jobs. Further we distinguish three arrival situations. The job arrives (1) during the switch-over time to queue 1, (2) during the service time of a class 1 job or (3) during a visit to queue 2. The probability for case (1) is s_i/C, for case (2) ρ_1 and for case (3) V_2/C. This results in

$$W_1 = Q_1 b_1 + \rho_1 R_1 + \frac{s_1}{C} R_{s_1} + \frac{V_2}{C}\left(R_{V_2} + s_i\right) , \tag{13.16}$$

where R_{V_2} denotes the mean residual visit time to queue 2. Substituting Little's law, $Q_1 = \lambda_1 W_1$, we get

$$Q_1 = \frac{\lambda_1}{1-\rho_1}\left(\rho_1 R_1 + \frac{s_1}{C} R_{s_1} + \frac{V_2}{C}\left(R_{V_2} + s_i\right)\right) . \tag{13.17}$$

In here there are still two unknowns, namely Q_1 and R_{V_2}. In order to obtain more equations we first condition Q_1 on the visit time. With $Q_{i,j}$ the mean number of class i jobs during a visit to queue j, we have

$$Q_1 = \frac{V_1}{C} Q_{1,1} + \frac{V_2}{C} Q_{1,2} . \tag{13.18}$$

Note that, because of the exhaustive discipline, there are no class 1 jobs left at the start of the visit to queue 2. So $Q_{1,2}$ is just the mean number of class 1 jobs that arrived during the visit to queue 2. Since the mean age of the visit to queue 2 is equal to the mean residual visit time to queue 2, it follows

$$Q_{1,2} = \lambda_1 R_{V_2} . \tag{13.19}$$

Finally, we observe that, because of the exhaustive discipline,

$$R_{V_2} = Q_{2,2} \frac{b_2}{1-\rho_2} + \frac{s_2}{V_2} \frac{R_{s_2}}{1-\rho_2} + \frac{\rho_2 C}{V_2} \frac{R_2}{1-\rho_2} . \tag{13.20}$$

Similarly, starting from a class 2 arrival, one gets another set of equations. From the equations (13.17)-(13.20) and the corresponding ones for a class 2 arrival, the unknowns $Q_{i,j}$, the two Q_i and the two R_{V_i} can be readily solved. So we see that, although the reasoning has been a little bit less standard, still only MVA arguments are needed to obtain the mean waiting times for this polling model.

13.4 AMVA for single-station systems

When an exact analysis is intractable, the ideas of MVA may also be applied heuristically to obtain approximations; this is demonstrated below for some single-station systems.

13.4.1 M|G|c

In the $M|G|c$ system, no exact results are available for the mean waiting time W, but the MVA approach can be used heuristically to derive an excellent approximation. Let λ be the arrival rate, b the mean service time and R the mean residual service time, such that $\rho = \lambda b/c < 1$ (to guarantee stability). If all servers are busy on arrival, the job first has to wait until the first departure and then continues to wait for as many departures as there were jobs waiting upon arrival. By assuming, as an approximation, that departures occur c times faster with c servers than with a single server, we get (cf. (13.13))

$$W = B\frac{R}{c} + Q\frac{b}{c}, \tag{13.1}$$

where B is the probability that all servers are busy. Although this probability is not known exactly, it appears to be fairly insensitive to the service time distribution. Hence, it can be well approximated by (13.12), the probability that all servers are busy in the $M|M|c$ with the same arrival rate and same mean service times. Another heuristic derivation of (13.1) is the following. Conditioning on the event that all servers are busy on arrival, and assuming that R is the mean residual service time of each server, the mean total amount of work just before the arrival is $cR + Q_+ b$, where Q_+ is the mean conditional queue length. By assuming that the mean residual service times are also R when the job goes into service, the mean total amount of work just before the arrival is also equal to $(c-1)R + cW_+$, where W_+ is the mean conditional waiting time. Hence,

$$cR + Q_+ b = (c-1)R + cW_+.$$

Multiplying with B, and using $BQ_+ = Q$ and $BW_+ = W$, this equation reduces to (13.1). From (13.1), together with Little's law (13.1), we obtain the approximation

$$W = \frac{B}{1-\rho}\,\frac{R}{c},$$

where B is given by (13.12).

13.4.2 M|G|c *with priorities*

The approximation for the mean waiting of the previous model can be readily extended to the $M|G|c$ non-preemptive priority system treating C classes of jobs, all with statistically equal service times. Again by assuming, as an approximation, that departures occur c times faster with c servers than with a single server, we get, similar to (13.5) in the single-server case,

$$W_i = B\frac{R}{c} + \sum_{j\leq i} Q_j \frac{b}{c} + \sum_{j<i} \lambda_j W_i \frac{b}{c} \tag{13.2}$$

and thus also, similar to (13.9),

$$W_i = \frac{B}{(1-\sum_{j\leq i}\rho_j)(1-\sum_{j<i}\rho_j)}\,\frac{R}{c}, \qquad i = 1,\ldots,C.$$

13.5 ASTA property in PF networks

We now move our attention from single station systems to networks of stations. For ease of presentation we will restrict our attention to multi-server stations servicing jobs according to the FCFS discipline. The basic tools of MVA for single station systems are the PASTA property and Little's law. Clearly, the latter is also applicable to networks of stations, but the PASTA property does not hold any longer, as there need not be Poisson processes inside the network. Hence, in this section, we will establish an extension of PASTA to PF networks, i.e., the so-called ASTA property or *Arrival Theorem.*

13.5.1 *Open single-class PF networks*

We first consider an open single-class network consisting of M multi-server stations, numbered $m = 1,2,\ldots,M$, and each with c_m exponential servers with service rate μ_m. Jobs arrive at the network according to a Poisson process at rate λ and enter the network in station m with probability q_m. The routing of jobs through the network is Markovian: after visiting station m, a job moves to station n with probability p_{mn} or leaves the system (with probability p_{m0} (so $\sum_{n=0}^{M} p_{mn} = 1$). Let P be the matrix of routing probabilities p_{mn}, $m,n = 1,\ldots,M$. We assume that P^n tends to 0 as n tends to infinity, which means that each job will eventually leave the network. Let Λ_m denote the total (external and internal) arrival rate at station m. These rates are the unique solution of the so-called traffic equations,

$$\Lambda_m = \lambda q_m + \sum_{n=1}^{M} \Lambda_n p_{nm}, \qquad m = 1,\dots,M. \tag{13.1}$$

For stability it is now assumed that $\Lambda_m < c_m\mu_m$ for $m = 1,\dots,M$. Due to the assumptions of exponential interarrival and service times and Markovian routing, the state of the network is fully described by the vector $\underline{k} = (k_1,\dots,k_M)$, where $k_m (\geq 0)$ denotes the number of jobs in station m. It is well-known (cf. Jackson [7]) that the equilibrium probability $p(\underline{k})$ has a product form, i.e.,

$$p(\underline{k}) = \frac{1}{G} p_1(k_1) p_2(k_2) \cdots p_M(k_M) \ , \tag{13.2}$$

where G is the normalization constant and $p_m(k_m)$ are (up to a constant) identical to the queue length probabilities of the $M|M|c_m$ with arrival rate Λ_m and service rate μ_m. So, writing $v_m(l) = \min(c_m, l)$,

$$p_m(k_m) = \prod_{l=1}^{k_m} \frac{1}{v_m(l)} \left(\frac{\Lambda_m}{\mu_m}\right)^{k_m} , \qquad k_m \geq 0 \ . \tag{13.3}$$

For this network the arrival theorem is more or less a copy of the PASTA property: *If a job enters the network in station m, jumps from station m to n or leaves the network in station m, it always sees the rest of the system in equilibrium.* The proof heavily relies on the PF equations (13.2)-(13.3). We demonstrate this for a job jumping from station m to n. The probability that this job sees the rest of the network in state $\underline{k}$ is equal to the number of times per time unit that a job jumps from m to n and sees $\underline{k}$ divided by the total number of jumps per time unit from m to n. By (13.2), the numerator is equal to

$$p(\underline{k} + \underline{e}_m) v_m(k_m + 1) \mu_m p_{mn} = p(\underline{k}) \Lambda_m p_{mn},$$

where $\underline{e}_m$ denotes the unit vector with a one at position m. Since the denominator is equal to $\Lambda_m p_{mn}$, it follows that the probability that the network is seen in state $\underline{k}$ is $p(\underline{k})$, which proves the arrival theorem.

13.5.2 Open multi-class PF networks

Let us consider a network servicing C different job classes. Class i jobs arrive according to a Poisson process at rate λ_i, $i = 1,\dots,C$. Per class the routing is again Markovian: class i jobs enter the network in station m with probability q_{im}, and after visiting station m, they move to station n with probability p_{imn} or leave the system with probability p_{im0}. The service rate of each of the c_m servers in station m is the same μ_m for all job classes. For stability it is needed that $\sum_{i=1}^{C} \Lambda_{im} < c_m\mu_m$ for $m = 1,\dots,M$, where Λ_{im} is the total arrival rate of class i at station m. These rates can be determined from traffic equations analogous to (13.1). The *global* state of the network can be described by the vector $\underline{k} = (\underline{k}_1,\ldots,\underline{k}_M)$, where $\underline{k}_m$ is again a vector, i.e., $\underline{k}_m = (k_{1m},\ldots,k_{Cm})$ where k_{im} denotes the number of class i jobs in station m. The total number of jobs in station m is denoted by k_m. Note that this global description does not lead to a Markov process. To obtain a Markov process, though, a detailed state description is required by including the order in which the jobs are waiting in the queue. In terms of the global states, the equilibrium probability $p(\underline{k})$ assumes the form

$$p(\underline{k}) = \frac{1}{G} p_1(\underline{k}_1) p_2(\underline{k}_2) \cdots p_M(\underline{k}_M) \,, \tag{13.4}$$

where

$$p_m(\underline{k}_m) = p_m(k_{1m}, k_{2m}, \ldots, k_{Cm}) = \left(\begin{array}{c} k_m \\ k_{1m},\ldots,k_{Cm} \end{array} \right) \prod_{l=1}^{k_m} \frac{1}{\nu_m(l)} \prod_{i=1}^{C} \left(\frac{\Lambda_{im}}{\mu_m} \right)^{k_{im}} . \tag{13.5}$$

The multinomial coefficient reflects the (remarkable!) property that, in terms of the detailed state description, all queue orders are equally likely. Based on the above PF, it can be shown, along the same lines as in the previous section, that a job arriving at a station sees the network in equilibrium.

13.5.3 Closed multi-class PF networks

In this section we consider a closed multi-class network. The number of class i jobs circulating in the network is K_i, $i = 1,\ldots,C$, where K_i is referred to as the population of class i and $\underline{K} = (K_1,\ldots,K_C)$ is the population vector. After visiting station m, class i jobs move to station n with probability p_{imn}, where $\sum_{n=1}^{M} p_{imn} = 1$, so jobs cannot escape from the network. Let f_{im} denote the *relative* arrival rate or *visiting frequency* of class i at station m. These rates satisfy the following traffic equations (cf. (13.1)),

$$f_{im} = \sum_{n=1}^{M} f_{in} p_{inm}, \qquad m = 1,\ldots,M.$$

Note that the 'real' arrival rates $\Lambda_{im}(\underline{K})$, which now depend on the population $\underline{K}$, are more difficult to obtain (see section 13.7), but they clearly satisfy

$$\frac{\Lambda_{im}(\underline{K})}{\Lambda_{in}(\underline{K})} = \frac{f_{im}}{f_{in}} \,. \tag{13.6}$$

The equilibrium probability $p(\underline{k})$ satisfies the same PF equations (13.4)-(13.5) as the open system, the only difference being that Λ_{im} is now replaced by f_{im} and the normalization constant G is computed as

$$G = G(\underline{K}) = \sum_{\underline{k} \in S(\underline{K})} p_1(\underline{k}_1) p_2(\underline{k}_2) \cdots p_M(\underline{k}_M) \,,$$

where $S(\underline{K})$ is the set of all global states in a network with population $\underline{K}$. The arrival theorem for a closed network with population $\underline{K}$ states that *if, after completing service, a class i job jumps from station m to n, it sees the rest of the network in the equilibrium that corresponds to a population $\underline{K}-\underline{e}_i$, i.e., with one job of its own class less*. This is a remarkable result: although the job we are looking at has always been in the system and thus has influenced the process in the past, when we look at the system at a jump moment, the rest of the system looks as if the job has never been there. The proof is based on the PF (13.4)-(13.5). The probability that a class i job jumping from station m to n sees the rest of the network in state $\underline{k}$ (in $S(\underline{K}-\underline{e}_i)$) is equal to the number of class i jumps per time unit from m to n seeing state $\underline{k}$ divided by the total number of class i jumps per time unit from m to n. By (13.4)-(13.5), the numerator is equal to

$$\begin{aligned} &p(\underline{k}_1,\ldots,\underline{k}_m+\underline{e}_i,\ldots,\underline{k}_M)\frac{k_{im}+1}{k_m+1}v_m(k_m+1)\mu_m p_{imn} \\ &= \frac{1}{G(\underline{K})}p_1(\underline{k}_1)\cdots p_m(\underline{k}_m)\cdots p_M(\underline{k}_M)f_{im}p_{imn}\ , \end{aligned}$$

where we use that any order of jobs in station m is equally likely, and thus the probability that a departure from m is of class i is equal to $(k_{im}+1)/(k_m+1)$. Accordingly, the denominator can be rewritten as

$$\begin{aligned} &\sum_{\underline{l}\in S(\underline{K}-\underline{e}_i)} p(\underline{l}_1,\ldots,\underline{l}_m+\underline{e}_i,\ldots,\underline{l}_M)\frac{l_{im}+1}{l_m+1}v_m(l_m+1)\mu_m p_{imn} \\ &= \frac{1}{G(\underline{K})}\sum_{\underline{l}\in S(\underline{K}-\underline{e}_i)} p_1(\underline{l}_1)\cdots p_M(\underline{l}_M)f_{im}p_{imn} = \frac{G(\underline{K}-\underline{e}_i)}{G(\underline{K})}f_{im}p_{imn}\ . \end{aligned}$$

Hence, the probability that the network is seen in state $\underline{k}$ is equal to

$$\frac{1}{G(\underline{K}-\underline{e}_i)}p_1(\underline{k}_1)\cdots p_M(\underline{k}_M)\ ,$$

which is the equilibrium probability for the network with population $\underline{K}-\underline{e}_i$ of being in state $\underline{k}$, and thus proves the arrival theorem. Finally, it is worthwhile to mention that the product-form solution, and thus also the arrival theorem for open and closed networks remain valid for *fixed* routing. Thus, for example, a closed network in which class 1 jobs cyclically visit stations 1, 2, 1, 3, 1, 2, ... (so $f_{11}=2, f_{12}=f_{13}=1$), also is of product form. Now we are equipped with the right tools, ASTA and Little's law, to apply MVA to networks of queues.

13.6 MVA for open PF networks

In this section we consider the open multi-class PF network introduced in Section 13.5.2. Let W_{im} denote the mean waiting time of class i jobs in station m and Q_{im} the mean number of class i jobs waiting in the queue. Further, B_m is the probability that all servers in station m are occupied, or by ASTA, the probability that an arriving job has to wait. Then, by ASTA, we get for the waiting time of an arriving class i job (cf. (13.13))

$$W_{im} = B_m \frac{1}{c_m\mu_m} + \sum_{j=1}^{C} Q_{jm} \frac{1}{c_m\mu_m} \; .$$

Clearly, the waiting time in station m does not depend on the class. Hence, together with Little's law, $Q_{jm} = \Lambda_{jm} W_{jm} = \Lambda_{jm} W_{im}$, this immediately yields

$$W_{im} = \frac{B_m}{1-\rho_m} \frac{1}{c_m\mu_m} \; , \qquad (13.1)$$

where ρ_m denotes the server utilization, given by

$$\rho_m = \frac{\sum_{j=1}^{C} \Lambda_{jm}}{c_m\mu_m} \; .$$

To determine B_m, note that from (13.4)-(13.5) one may easily show that the marginal distribution of the total number of jobs in station m is identical to the distribution of the number of jobs in the $M|M|c_m$ with arrival rate $\sum_{j=1}^{C} \Lambda_{jm}$ and service rate μ_m. Hence, B_m is equal to (cf. (13.12))

$$B_m = \frac{(c_m\rho_m)^{c_m}}{c_m!} \left((1-\rho_m) \sum_{n=0}^{c_m-1} \frac{(c_m\rho_m)^n}{n!} + \frac{(c_m\rho_m)^c}{c!} \right)^{-1} \; . \qquad (13.2)$$

13.7 MVA for closed single-class PF networks

The analysis of closed PF networks appears to be slightly more complicated than for open networks. It will be first demonstrated for the simplest case, namely:

Single class, single servers

Let us consider a closed single-class PF network with single server stations ($c_m = 1$ for all m) and population K. By ASTA, the mean waiting time $W_m(K)$ of an arriving job in station m satisfies

$$W_m(K) = \rho_m(K-1)\frac{1}{\mu_m} + Q_m(K-1)\frac{1}{\mu} \; . \tag{13.1}$$

Further we have Little's law,

$$Q_m(K) = \Lambda_m(K) W_m(K) \; , \tag{13.2}$$

and

$$\rho_m(K) = \Lambda_m(K)\frac{1}{\mu_m} \; , \tag{13.3}$$

The first difference with open networks is that equation (13.1) is recursive in the population size, and the second difference is that the arrival rate (or throughput) $\Lambda_m(K)$ is not known. Fortunately the rates $\Lambda_m(K)$ can be computed once the waiting times $W_m(K)$ are known. To see this, we add the equations (13.2)-(13.3) over all stations and use (13.6) to obtain

$$\begin{aligned} K &= \sum_{n=1}^{M} (Q_n(K) + \rho_n(K)) = \sum_{n=1}^{M} \Lambda_n(K)\left(W_n(K) + \frac{1}{\mu_n}\right) \\ &= \frac{\Lambda_m(K)}{f_m} \sum_{n=1}^{M} f_n \left(W_n(K) + \frac{1}{\mu_n}\right) \; . \end{aligned}$$

Hence,

$$\Lambda_m(K) = K\frac{f_m}{C(K)} \; , \tag{13.4}$$

where

$$C(K) = \sum_{n=1}^{M} f_n \left(W_n(K) + \frac{1}{\mu_n}\right) \; .$$

Note that $C(K)$ is the mean duration of a (generalized) cycle in which station n is visited f_n times, $n = 1, \ldots, N$. An appealing interpretation of (13.4) is the following. In each cycle, station m is visited f_m times, and thus $f_m/C(K)$ is the mean number of times per time unit that a (tagged) job is visiting station m. So, multiplication of $f_m/C(K)$ by the total number of circulating jobs obviously yields the throughput of station m. The relations (13.1)-(13.4) can be used to recursively calculate $W_m(k)$, $\Lambda_m(k)$, $Q_m(k)$ and $\rho_m(k)$ for populations starting from $k = 0$ up to $k = K$, where initially $Q_m(0) = \rho_m(0) = 0$ for all m.

Single class, multi servers

Let us now consider a network with multi-server stations. Then we only need to adapt the relation for $W_m(K)$, while (13.2) and (13.4) remain valid. In the multi-server case we have

$$W_m(K) = B_m(K-1)\frac{1}{c_m\mu_m} + Q_m(K-1)\frac{1}{c_m\mu_m} , \qquad (13.5)$$

where $B_m(K)$ denotes the probability that all servers in station m are occupied. Let $p_m(k;K)$ denote the probability of k jobs in station m, then

$$B_m(K) = \sum_{k=c_m}^{K} p_m(k;K) .$$

Thus, to compute $B_m(K)$, we need the marginal queue length probabilities of station m. These probabilities can be found by balancing the number of transitions per time unit between states $k-1$ and k: the rate from k to $k-1$ is $p_m(k;K)v_m(k)\mu_m$ and, by ASTA, the rate from $k-1$ to k is $\Lambda_m(K)p_m(k-1;K-1)$. Hence,

$$p_m(k;K) = \frac{\Lambda_m(K)}{v_m(k)\mu_m} p_m(k-1;K-1) , \qquad k = 1,\ldots,K, \qquad (13.6)$$

where $p_m(0;K)$ follows from the normalization,

$$p_m(0;K) = 1 - \sum_{k=1}^{K} p_m(k;K) . \qquad (13.7)$$

The relations (13.2) and (13.4)-(13.7) form an algorithm to recursively calculate $W_m(k)$, $\Lambda_m(k)$, $Q_m(k)$ and $p_m(\cdot;k)$ for populations starting from $k=0$ up to $k=K$, where initially $Q_m(0)=0$ and $p_m(0;0)=1$ for all m. We remark that the "1−" in (13.7) may cause numerical problems in bottleneck stations (i.e., stations with an extremely high load). A numerically stable (but more involved) solution, though, is presented in Casale [4].

Single class, queue-dependent servers

This algorithm can be readily extended to networks with queue-dependent servers; see Reiser [10]. The amount of work of each job in station m is assumed to be exponentially distributed with mean $1/\mu_m$ and the (single!) server works at rate $v_m(k)$ when k jobs are present in station m. Note that in the special case $v_m(k) = \min(c_m, k)$, station m reduces to an 'ordinary' multi-server station with c_m servers. For this network, relation (13.5) needs to be adapted, since the waiting time and actually, the whole sojourn time of an arriving job is not only determined by the situation encountered on arrival, but also by arrivals after the one of the tagged job. To obtain a relation for the mean sojourn time, we apply Little's law to station m, yielding

$$\Lambda_m(K)S_m(K) = L_m(K) = \sum_{k=1}^{K} kp_m(k;K) .$$

Dividing both sides of the above equation by $\Lambda_m(K)$ and using (13.6), we obtain

$$S_m(K) = \sum_{k=1}^{K} \frac{k}{v_m(k)\mu_m} \, p_m(k-1; K-1) \; .$$

Together with Little's law, $L_m(K) = \Lambda_m(K) S_m(K)$, the equation for the throughput

$$\Lambda_m(K) = K \frac{f_m}{\sum_{n=1}^{M} f_n S_n(K)} \; ,$$

and the equations (13.6)-(13.7), we can again recursively calculate $S_m(k)$, $\Lambda_m(k)$, $L_m(k)$ and $p_m(\cdot; k)$ for all populations $k = 0$ up to K.

13.8 MVA for closed multi-class PF networks

The extension to closed multi-class networks with multi-server stations is straightforward. The mean waiting time $W_{im}(\underline{K})$ of an arriving class i job in station m satisfies

$$W_{im}(\underline{K}) = B_m(\underline{K} - \underline{e}_i) \frac{1}{c_m \mu_m} + \sum_{j=1}^{C} Q_{jm}(\underline{K} - \underline{e}_i) \frac{1}{c_m \mu_m} \; . \tag{13.1}$$

Note that, as opposed to open multi-class networks, the mean waiting time does depend on the class. Further, we have

$$\Lambda_{im}(\underline{K}) = K_i \, \frac{f_{im}}{\sum_{n=1}^{M} f_{in} \left(W_{in}(\underline{K}) + \frac{1}{\mu_n} \right)} \; . \tag{13.2}$$

and by Little's law,

$$Q_{im}(\underline{K}) = \Lambda_{im}(\underline{K}) W_m(\underline{K}) \; . \tag{13.3}$$

The probability $B_m(\underline{K})$ of c_m busy servers in station m is equal to

$$B_m(\underline{K}) = \sum_{k=c_m}^{K_1+\cdots+K_C} p_m(k; \underline{K}) \; ,$$

where the queue length probabilities $p_m(k; \underline{K})$ satisfy

$$p_m(k; \underline{K}) = \sum_{j=1}^{C} \frac{\Lambda_{jm}(\underline{K})}{v_m(k)\mu_m} p_m(k-1; \underline{K} - \underline{e}_j) \; , \qquad k = 1, \ldots, K_1 + \cdots + K_C, \tag{13.4}$$

and $p_m(0; \underline{K})$ follows again from the normalization equation. Together, relations (13.1)-(13.4) form a recursive algorithm for the calculation of the mean waiting times in a closed network. Note that this algorithm suffers from the curse of dimensionality: the complexity of the algorithm is determined by the number of steps needed to go through all $\prod_{i=1}^{C}(K_i+1)$ sub-populations of $\underline{K}$, and thus, already for a moderate number of classes and number of jobs per class, the number of steps becomes so large that it is no longer feasible to execute this algorithm. Fortunately, as we will see in section 13.12, there exist good approximations.

13.9 AMVA for open networks

Exact analysis of open networks with generally distributed service times is extremely difficult and in most cases intractable. For large randomly routed networks, however, it is reasonable to expect that heuristic application of MVA might produce fairly accurate predictions for mean waiting times (see also chapters 5 and 6 in Buzacott and Shanthikumar [3]). To illustrate this, we consider an open multi-class network with multi-server stations, and denote by b_{im} and R_{im} the mean service time and mean residual service time, respectively, of class i jobs in station m. Then the mean service time b_m and mean residual service time R_m of an *arbitrary* job in station m are

$$b_m = \sum_{i=1}^{C} \frac{\Lambda_{im}}{\Lambda_m} b_{im}, \qquad R_m = \sum_{i=1}^{C} \frac{\rho_{im}}{\rho_m} R_{im},$$

where $\Lambda_m = \sum_{i=1}^{C} \Lambda_{im}$, $\rho_{im} = \Lambda_{im} b_{im}/c_m$ and $\rho_m = \sum_{i=1}^{C} \rho_{im}$. Then for all classes the mean waiting time in station m approximately satisfies (cf. (13.1))

$$W_m = B_m \frac{R_m}{c_m} + Q_m \frac{b_m}{c_m},$$

where B_m is approximated by (13.2). Together with Little's law, $Q_m = \Lambda_m W_m$, this immediately yields the approximation

$$W_m = \frac{B_m}{1-\rho_m} \frac{R_m}{c_m}.$$

The mean sojourn time in station m does depend on the class; for a class i job we have $S_{im} = W_m + b_{im}$. The above approximation may work well for large randomly routed networks, but in other cases it might be better to resort to decomposition approaches that approximate each station by a $G|G|c$ queue with an appropriate arrival process, see e.g. Whitt [18].

13.10 AMVA for closed single-server networks

In this section we consider closed single-server networks. The case of multi-server networks is briefly discussed in the next section. So far we only considered PF networks. Now we want to relax some of the conditions required for the existence of a PF. The first one concerns the requirement of the same exponential service times for all job classes. The second one is the FCFS service discipline. Relaxing these conditions leads to non-PF networks, and thus we have to look for approximations.

13.10.1 Class-dependent general service times

Let us assume that the service time requirement of a class i job in station m follows a general distribution with mean b_{im} and let R_{im} denote the mean residual service time. Heuristic application of ASTA leads to

$$W_{im}(\underline{K}) = \sum_{j=1}^{C} (\rho_{jm}(\underline{K} - \underline{e}_i)R_{jm} + Q_{jm}(\underline{K} - \underline{e}_i)b_{jm}), \tag{13.1}$$

where $\rho_{jm}(\underline{K})$ is the utilization of station m by class j jobs, i.e., $\rho_{jm}(\underline{K}) = \Lambda_{jm}(\underline{K})b_{jm}$. The relation for the throughput $\Lambda_{im}(\underline{K})$ is given by (cf. (13.2))

$$\Lambda_{im}(\underline{K}) = K_i \, \frac{f_{im}}{\sum_{n=1}^{M} f_{in} \, (W_{in}(\underline{K}) + b_{in})} \, .$$

and by Little's law, the mean number in the queue satisfies $Q_{im}(\underline{K}) = \Lambda_{im}(\underline{K})W_{im}(\underline{K})$.

The above relations provide an MVA algorithm for the calculation of mean waiting times. However, the errors can be significant when the variability of the service times is large. To easily see this, consider the single-class case and write (see (13.3)),

$$R_m = \frac{b_m}{2}(1 + c_m^2),$$

where c_m^2 is the squared coefficient of variation of the service time. Then (13.1) simplifies to

$$W_m(K) = \rho_m(K-1) \, \frac{b_m}{2}(1 + c_m^2) + Q_m(K-1)b_m \, .$$

Multiplication of both sides by $\Lambda_m(K)$, yields

$$Q_m(K) = \rho_m(K-1)\frac{\rho_m(K)}{2}(1 + c_m^2) + Q_m(K-1)\rho_m(K) \, .$$

Clearly, the left-hand side is bounded by K, while the right-hand side can get arbitrarily large as c_m^2 tends to infinity. Hence, in using AMVA relations such as (13.1), one should be careful when the variability of the service times is very large.

13.10.2 Priorities

Now consider the case jobs in station m are no longer treated FCFS, but according to non-preemptive priorities, where class 1 has the highest and class C the lowest priority. Then, by heuristic application of ASTA, relation (13.1) should be adapted as

$$\begin{aligned} W_{im}(\underline{K}) &= \sum_{j=1}^{C} \rho_{jm}(\underline{K}-\underline{e}_i)R_{jm} + \sum_{j\leq i} Q_{jm}(\underline{K}-\underline{e}_i)b_{jm} + W_{im}(\underline{K}) \sum_{j<i} \Lambda_{jm}(\underline{K}-\underline{e}_i)b_{jm} \\ &= \sum_{j=1}^{C} \rho_{jm}(\underline{K}-\underline{e}_i)R_{jm} + \sum_{j\leq i} Q_{jm}(\underline{K}-\underline{e}_i)b_{jm} + W_{im}(\underline{K}) \sum_{j<i} \rho_{jm}(\underline{K}-\underline{e}_i). \end{aligned} \quad (13.2)$$

The relations for $\Lambda_{im}(\underline{K})$ and $Q_{im}(\underline{K})$ remain the same.

13.10.3 Multiple visits to a station

Consider a network where each class of jobs can make several visits to a station during a cycle, each visit involving a different exponential service requirement. This is again not a PF, but it may be modeled by AMVA as follows. Let n_{im} denote the number of distinct types of service that a class i job receives at station m and let $1/\mu_{imk}$ denote the mean service requirement for a type k service. Further, let f_{imk} denote the mean number visits to station m during a cycle requiring a type k service. Then the mean sojourn time of a class i job in station m receiving a type k service is (approximately) equal to

$$S_{imk}(\underline{K}) = \sum_{j=1}^{C} \sum_{l=1}^{n_{jm}} L_{jml}(\underline{K}-\underline{e}_i) \, \frac{1}{\mu_{jml}} + \frac{1}{\mu_{imk}},$$

where $L_{jml}(\underline{K})$ is the mean number of class j jobs in station m for a type l service. By Little's law,

$$L_{imk}(\underline{K}) = \Lambda_{imk}(\underline{K})S_{imk}(\underline{K}),$$

and the throughput $\Lambda_{imk}(\underline{K})$ satisfies

$$\Lambda_{imk}(\underline{K}) = K_i \, \frac{f_{imk}}{\sum_{n=1}^{M} \sum_{l=1}^{n_{in}} f_{inl} S_{inl}(\underline{K})} \, .$$

The above set of equations leads again to a recursive algorithm to compute mean sojourn times and mean cycle times.

13.11 AMVA for closed multi-server networks

In this section we briefly discuss AMVA for multi-server networks. For ease of presentation we assume that the distribution of the service requirement in station m is the same for each class, with mean b_m. Let R_m denote the mean residual service requirement. AMVA yields (cf. (13.1))

$$W_{im}(\underline{K}) = B_m(\underline{K}-\underline{e}_i)\frac{R_m}{c_m} + \sum_{j=1}^{C} Q_{jm}(\underline{K}-\underline{e}_i)\frac{b_m}{c_m} \ .$$

Further,

$$\Lambda_{im}(\underline{K}) = K_i \ \frac{f_{im}}{\sum_{n=1}^{M} f_{in}\,(W_{in}(\underline{K})+b_n)} \ .$$

and by Little's law, $Q_{im}(\underline{K}) = \Lambda_{im}(\underline{K})W_{im}(\underline{K})$.

To estimate $B_m(\underline{K})$ we may avoid the calculation of queue length probabilities by approximating $B_m(\underline{K})$ by the probability of c_m busy severs in the corresponding $M|M|c_m$ queue with arrival rate $\sum_{j=1}^{C}\Lambda_{jm}(\underline{K})$ and mean service time b_m,

$$B_m(\underline{K}) = \frac{(c_m\rho_m(\underline{K}))^{c_m}}{c_m!}\left((1-\rho_m(\underline{K}))\sum_{n=0}^{c_m-1}\frac{(c_m\rho_m(\underline{K}))^n}{n!} + \frac{(c_m\rho_m(\underline{K}))^c}{c!}\right)^{-1} \ ,$$

where $\rho_m(\underline{K}) = \sum_{j=1}^{C}\Lambda_{jm}(\underline{K})b_m/c_m$. This completes again the set of equations that can be used to recursively calculate the mean waiting times. The extension from FCFS to priorities is straightforward. If jobs in station m are served according to non-preemptive priorities, with class 1 given the highest priority, then the relation for the mean waiting time becomes (cf. (13.2) and (13.2))

$$\begin{aligned} W_{im}(\underline{K}) &= B_m(\underline{K}-\underline{e}_i)\frac{R_m}{c_m} + \sum_{j\le i} Q_{jm}(\underline{K}-\underline{e}_i)\frac{b_m}{c_m} + W_{im}(\underline{K})\sum_{j<i}\Lambda_{jm}(\underline{K}-\underline{e}_i)\frac{b_m}{c_m} \\ &= B_m(\underline{K}-\underline{e}_i)\frac{R_m}{c_m} + \sum_{j\le i} Q_{jm}(\underline{K}-\underline{e}_i)\frac{b_m}{c_m} + W_{im}(\underline{K})\sum_{j<i}\rho_{jm}(\underline{K}-\underline{e}_i) \ , \end{aligned}$$

where $\rho_{jm}(\underline{K}) = \Lambda_{jm}(\underline{K})b_m/c_m$.

13.12 The Schweitzer-Bard approximation

The MVA algorithm for closed multi-class networks suffers from the curse of dimensionality. To illustrate the problem and one of its solutions, we consider a multi-class PF network with single-server FCFS stations and class-independent exponential service times. The mean sojourn time $S_{im}(\underline{K})$ of a class i job in station m satisfies

$$S_{im}(\underline{K}) = \sum_{j=1}^{C} L_{jm}(\underline{K} - \underline{e}_i)\frac{1}{\mu_m} + \frac{1}{\mu_m} , \tag{13.1}$$

where by Little's law, the mean number of jobs $L_{im}(\underline{K})$ in station m is given by

$$L_{im}(\underline{K}) = \Lambda_{im}(\underline{K})S_{im}(\underline{K}) , \tag{13.2}$$

and the class i throughput of station m follows from

$$\Lambda_{im}(\underline{K}) = K_i \frac{f_{im}}{\sum_{n=1}^{M} f_{in}S_{in}(\underline{K})} . \tag{13.3}$$

The set of equations (13.1)-(13.3) forms a recursive algorithm to compute the $3CM$ mean values $S_{im}(\underline{K})$, $L_{im}(\underline{K})$ and $\Lambda_{im}(\underline{K})$. However, the number of sub-populations of $\underline{K}$ one needs to go through is $\prod_{i=1}^{C}(K_i + 1)$. Hence, already for a moderate number of classes and number of jobs per class, the number of sub-populations becomes too large. Then we need an approximation. One such an approximation is due to Schweitzer [11] and Bard [2]. The idea is to break the recursion in (13.1) by adopting the following approximation assumption: *an arriving type i job sees the system in equilibrium with a population* $\underline{K}$ *instead of* $\underline{K} - \underline{e}_i$. Thus the mean number of jobs seen on arrival is the mean number in a network *including himself.* But, of course, the arriving class i job does not have to wait for himself. Therefore, to avoid self queueing, the mean number $L_{im}(\underline{K})$ is multiplied by the factor $(K_i - 1)/K_i$ (which vanishes when $K_i = 1$; see also [12]). Hence, it is assumed that, approximately,

$$L_{jm}(\underline{K} - \underline{e}_i) = L_{jm}(\underline{K}), \qquad j \neq i, \tag{13.4}$$

and

$$L_{im}(\underline{K} - \underline{e}_i) = \frac{K_i - 1}{K_i} L_{im}(\underline{K}). \tag{13.5}$$

Substitution of (13.4)-(13.5) in (13.1) results in

$$S_{im}(\underline{K}) = \sum_{j \neq i} L_{jm}(\underline{K})\frac{1}{\mu_m} + \frac{K_i - 1}{K_i} L_{im}(\underline{K}) + \frac{1}{\mu_m} . \tag{13.6}$$

Hence, the recursive set of equations (13.1)-(13.3) is turned into a set of fixed point equations (13.2), (13.3) and (13.6) for $3CM$ unknowns, namely $S_{im}(\underline{K})$, $L_{im}(\underline{K})$ and $\Lambda_{im}(\underline{K})$. Its solution can be found by successive substitutions. In practice, successive substitutions converges quickly. In theory, however, convergence and uniqueness of the solution of the set of equations (13.2), (13.3) and (13.6) is still an open problem. Typically, the result of the Schweitzer-Bard approximation is within $5 - 10\%$ of the exact values for the throughputs $\Lambda_{im}(\underline{K})$ and within $15 - 30\%$ of the exact values for the mean values $S_{im}(\underline{K})$ and $L_{im}(\underline{K})$.

There are quite a few ways in which one can approximately solve the MVA equations. In [5] several of these approaches are formulated in a unifying framework.

A simple improvement on this fixed point scheme is to combine MVA and the

Schweitzer-Bard approximation as follows; see, e.g., Chapter 4 in Van Doremalen [6]. First use the Schweitzer-Bard fixed point scheme to approximate the performance characteristics for all possible population vectors with one job less, so $\underline{K} - \underline{e}_1, \ldots, \underline{K} - \underline{e}_C$. Then compute the performance characteristics for the population $\underline{K}$ in one MVA step. This approach is known as *first order depth improvement.* It reduces the errors to 1% with an occasional error of 5%. A further improvement is that the Schweitzer-Bard fixed point scheme is used for all populations with two jobs less and MVA thereafter. Then the errors become negligible.

Acknowledgement

This paper is dedicated to the memory of Paul J. Schweitzer, who, before we started writing, had agreed to be one of the authors of this paper. Although there are more names associated to the MVA approach, it seems right to remember Paul as the inventor of approximate MVA.

References

1. Artalejo, J.R. (1999). A classified bibliography of research on retrial queues: Progress in 1990-1999. *TOP* 7, pp. 187-211.
2. Bard, Y. (1979). Some Extensions to Multiclass Queueing Network Analysis. In: Performance of Computer Systems, Proceedings of the Third International Symposium on Modelling and Performance Evaluation of Computer Systems, Vienna, Austria, M. Arató, A. Butrimenko and Erol Gelenbe (eds.), North-Holland, Amsterdam, pp. 51-62.
3. Buzacott, J.A., Shanthikumar, J.G. Stochastic models of manufacturing systems. Prentice Hall, 1993.
4. Casale, G. (2008). A note on stable flow-equivalent aggregation in closed networks. *Queueing Syst. Theory Appl.*, **60**, pp. 193202.
5. Cremonesi, P., Schweitzer, P.J. and Serrazi, G. (2002). A unifying framework for the approximate solution of closed multiclass queueng networks. *IEEE Transactions on Computers*. **51**, pp.1423-1434.
6. Doremalen, J.B.M. van (1986). Approximate analysis of queueing network models. *PhD Thesis*, Eindhoven University of Technology.
7. Jackson, J.R. (1957). Networks of waiting lines. *Oper. Res.* **5**, pp. 518-521.
8. Lavenberg, S.S., Reiser, M. (1980). Stationary state probabilities at arrival instants for closed queueing networks with multiple types of customers. *J. Appl. Prob.* **17**, pp. 1048-1061.
9. Reiser, M., Lavenberg, S.S. (1980). Mean value analysis of closed queueing networks. *Journal of the ACM* **27**, pp. 313-322. *Corrigendum Journal of the ACM* **28**, pg. 629.
10. Reiser, M. (1981). Mean value analysis and convolution method for queue-dependent servers in closed queueing networks. *Perf. Eval.* **1**, pp. 7-18.
11. Schweitzer, P.J. (1979). Approximate analysis of multiclass closed networks of queues. *Proceedings of the International Conference on Stochastic Control and Optimization*, Amsterdam.
12. P.J. Schweitzer, P.J., Seidmann, A. and Shalev-Oren, S. (1986). The correction terms in Approximate Mean Value Analysis. *Opns. Res. Letters* **4**, pp. 197-200.
13. Stidham, S. (1974). A last word on $L = \lambda W$. *Opns. Res.* **22**, pp. 417-421.
14. Wolff, R.W. (1982). Poisson arrivals see time averages. *Opns. Res.* **30**, pp. 223-231.

15. Kleinrock, L. Queueing systems, volume I: theory. John Wiley & Sons, 1975.
16. J.D. LITTLE, *A proof of the queueing formula $L = \lambda W$*, Opns. Res., 9 (1961), pp. 383–387.
17. S.M. ROSS, *Introduction to probability models*, 8th ed., Academic Press, London, 2002.
18. Whit, W. (1983). The queueing network analyzer. *Bell System technical Journal* **62**, pp. 2779-2815.
19. Yaskov, S. (1987). Processor-sharing queues: some progress in analysis. *Queueing Syst. Theory Appl.* **2**, pp. 1-17.

Chapter 14
Response Time Distributions in Networks of Queues

Michael Grottke, Varsha Apte, Kishor S. Trivedi, and Steve Woolet

Abstract This chapter addresses the issue of determining the response time distribution in networks of queues. Four different techniques are described and demonstrated. A two-step numerical approach to compute the response time distribution for closed Markovian networks with general connectivity, a technique for determining the approximate (exact under certain conditions) response time distribution of a defined subset of open $M/M/c/b$ Markovian networks using predefined continuous-time Markov chain (CTMC) "response time blocks," an expansion of "response time blocks" to open Markovian networks with general phase-type (PH) service time distributions, and an approach for handling non-Markovian networks having $M/G/1$ priority and $PH/G/1$ queues. These techniques are shown to give accurate results with much smaller CTMCs or semi-Markov processes than exact analysis.

14.1 Introduction

The problem of computing the response (sojourn) time distribution in queuing networks has been researched extensively during the past few decades. (For a somewhat dated survey see [6].) In case of open queuing networks, a considerable

Michael Grottke
Department of Statistics & Econometrics, University of Erlangen-Nuremberg, Germany, e-mail: Michael.Grottke@wiso.uni-erlangen.de
Varsha Apte
Department of Computer Science, IIT Bombay, India
e-mail: varsha@cse.iitb.ac.in
Kishor S. Trivedi
Department of Electrical & Computer Engineering, Duke University, Durham, USA
e-mail: kst@ee.duke.edu
Steve Woolet
IBM Systems & Technology Group, IBM Corporation, Durham, USA
e-mail: steve_woolet@us.ibm.com

R.J. Boucherie and N.M. van Dijk (eds.), *Queueing Networks: A Fundamental Approach*,
International Series in Operations Research & Management Science 154,
DOI 10.1007/978-1-4419-6472-4_14, © Springer Science+Business Media, LLC 2011

amount of work has been done in computing the response time distribution in the domain of Jackson networks. Closed form solutions have been derived for the (Laplace-Stieltjes transform of) response time distributions through a particular path in product-form queuing networks [38].

Furthermore, many results exist for response time distributions in networks with a specific topological structure such as tandem, central server, single queue with feedback and so on. The communications literature also shows a focus on end-to-end packet delay in tandem-type queuing networks, with characteristics specific to communication systems. However, it is very difficult to derive exact closed form solutions for networks with even slightly non-restrictive topology as well as service and arrival characteristics. In the face of such difficulties, the two approaches taken are: (1) numerical solution and (2) approximate solution.

In case of closed Markovian queuing networks, the *tagged customer approach* [30] may be used to numerically compute the response time distribution of a network with a general topology. However, this technique consists of generating the state space of the queuing network and may result in a very large state space. Section 14.2 will describe this numerical approach. Methods for efficiently analyzing response time densities in very large Markov and semi-Markov models have recently been shown in [7] and [12].

In case of open queuing networks with unlimited capacity queues, the tagged customer approach is not feasible at all. Thus, a lot of research has been devoted to finding approximations to response time distribution which are space and time efficient. Abate et al. [1] consider approximations for $G/GI/1$ queue sojourn time tail probabilities. Au-Yeng et al. [2] present a technique using generalized lambda distributions for approximating response time densities in Markov and semi-Markov models. Van Houdt and Blondia [44] approximate waiting time distributions by steady-state analysis of *reset* Markov chains. Van Velthoven et al. [45] present methods for calculating the response time distribution of impatient customers in discrete-time queues.

One class of approximations which address the response time distribution problem have been termed “independent flow time approximation” (IFTA) by Boxma and Daduna [6]. This approximation states that “arrival state distributions seen by a test customer on the arrival at successive network stations are independent of each other and equal the stationary arrival state distributions at these stations as seen by an arbitrary customer” [6]. It has been applied by Harrison [14], as well as Shanthikumar and Buzacott [40].

In the approach suggested by Harrison [14], decomposition of queues is used to compute the response time distribution. The basic idea is to find arrival rates to each queue and then analyze each queue in isolation. If we know the response time distribution of a job at each of the queues in the network, then the response time of a job through a particular path in the queuing network can be computed as the convolution of the response time distributions at each queue in the path.

A response time computation technique that builds on Harrison’s method was presented in its nascent form by Woolet [48]. The key idea there was as follows: Assume that the response time distribution at each queue in the queuing network is phase-type, i.e., it corresponds to the absorption time distribution of a continuous-time Markov chain (CTMC). Then if we construct the CTMCs corresponding to all queues, "glue" them together according to the topology of the network, and add an absorbing state representing departure from the network, we have another CTMC. The absorption time distribution of this CTMC approximates the response time distribution in the same way as Harrison's method does. The advantage of this method is that it (1) provides a clear representation, in the form of a CTMC, for the approximate computation technique, (2) maps the problem to a well-explored problem of transient solution of CTMCs and (3) allows us to extend the original technique to more general topologies and service time distributions, by taking advantage of this CTMC representation. This method also maintains the advantage of space and time efficiency that Harrison's method has. It must be noted that though we do use CTMCs in this method, they do not represent the state space of the queuing network. In fact, the size of the CTMC is linear in the number of queues in the network. In [48], this method was presented with some examples for a network of $M/M/c/b \leq \infty$ queues with no loss of customers. Section 14.3 will describe this application to open queuing networks. We extend this method in Sect. 14.4 to cover queues in which the service time follows a phase-type distribution.

Two approaches that can be adopted if the response time distribution at each queue is not phase-type are described in Sect. 14.5. First, we can fit a phase-type distribution to the response time distribution and again map the problem to a CTMC transient solution problem. Most of the results in the literature provide the response time distribution at a queue only in the form of its Laplace-Stieltjes transform (LST). This poses the additional problem of matching two distributions, given only their LSTs. We explore this problem and present our observations and suggestions.

Another approach is to leave the CTMC domain altogether and map the queuing network to a semi-Markov process as was done by Mainkar [28]. Since the response time distribution at each queue and the routing probabilities are available, the same idea of "gluing" together states which represent response time distributions will result in a semi-Markov process whose kernel, holding times in states, and embedded discrete-time Markov chain (DTMC) are known. Again, we approximate the response time distribution of the queuing network by the absorption time distribution of the semi-Markov process. The number of states of this semi-Markov process is linear in the number of queues in the queuing network.

In this chapter, these basic ideas are used to develop approximations of response time distributions for a variety of queuing networks.

14.2 Closed Markovian networks

As stated previously, closed form solutions for the response time distributions of queuing networks are obtainable only for simple queuing system models. Methods for computing the Laplace transforms of the response time distribution in queuing networks satisfying the non-overtaking condition are proposed by Lemoine [24] as well as Walrand and Varaiya [46]. Melamed and Yadin [31] have shown that numerical computation of the response time distribution is possible, although this requires the construction and solution of large Markov models.

14.2.1 Tagged customer approach

The tagged customer approach is one method of numerical computation of the response time distribution for queuing networks with general interconnectivity. Melamed and Yadin [31] present this approach for evaluating the response time distribution in a discrete state Markovian queuing network. An arbitrary customer is picked as the tagged customer and its passage is tracked through the network. The response time distribution, conditioned on the state of the queuing network at the time of arrival of the tagged customer, is computed. Deriving this conditional response time distribution of the tagged customer is transformed into a problem of solving for the distribution of the time to absorption of a finite-state CTMC. For closed product form queuing networks, an arriving customer will see the network in equilibrium with one less customer and we can establish the distribution of the other customers in the network at the instant of arrival of the tagged customer; see the arrival theorem of Sevcik and Mitrani [39] or Lavenberg and Reiser [23]. All the states in which the tagged customer may find the queuing network upon arrival (i.e., how the remaining jobs are distributed among the queues in the network) must be determined to make possible the computation of the unconditional response time distribution. Computing the response time distribution using the tagged customer approach is therefore a two-step process.

14.2.2 Example: Central server model

Figure 14.1 shows a closed queuing network model of a computer system, a central server model (CSM), that will be used to describe the tagged customer approach. Customers first join the CPU queue. The CPU, Disk 1 and Disk 2 are assumed to have exponentially distributed service times with parameters μ_C, μ_{D1} and μ_{D2}, respectively. The service discipline at all the queuing centers is assumed to be first come, first served (FCFS). This closed system contains n customers. A customer will request access to Disk 1 with probability p_1 and Disk 2 with probability p_2 after receiving a burst of service at the CPU; the customer rejoins the CPU queue for another burst of service after completing access to the disks. With probability $p_0 = 1-(p_1+p_2)$ the customer may complete execution, after which it is replaced by a statistically identical customer newly arriving to the system.

As mentioned earlier, finding the response time distribution of a queuing network is transformed into solving for the absorption time distribution of a CTMC. Fig-

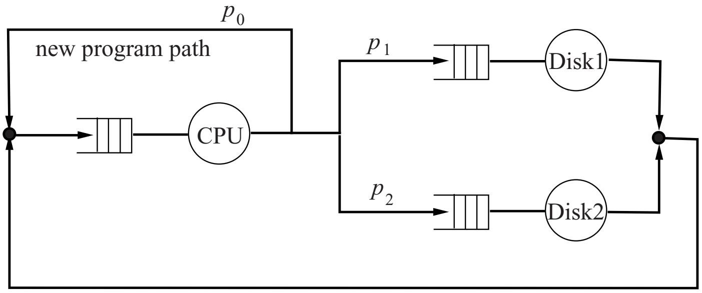


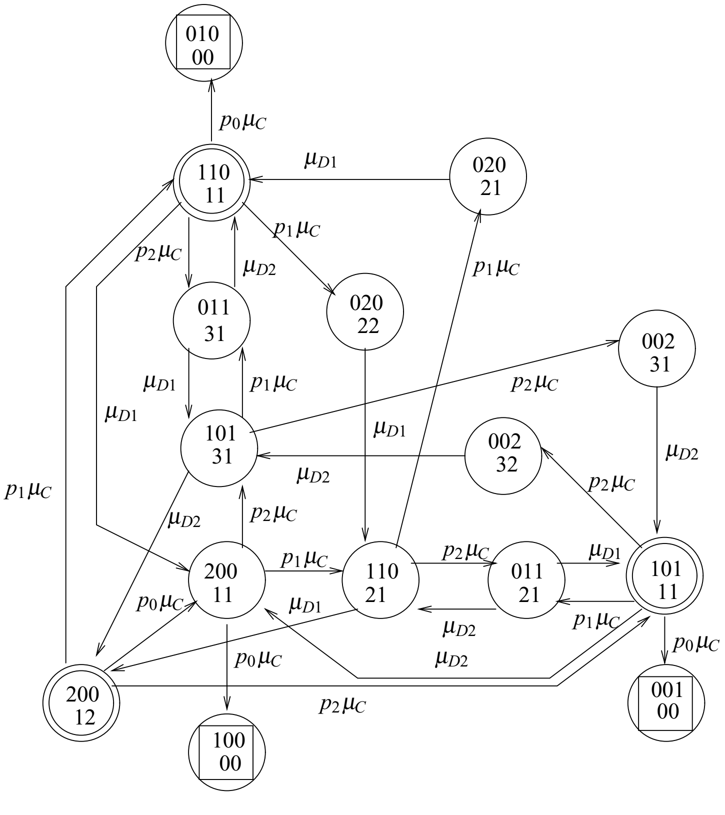
Fig. 14.1: Central server model of a computer system

ure 14.2 shows the CTMC whose absorption time distribution needs to be computed for the case of the CSM with only $n = 2$ customers.

It is interesting to note that a rather complex CTMC is obtained even for a system with only two customers. In Fig. 14.2, the number of customers at the CPU, Disk1, and Disk2, respectively, are indicated by the first three components of the state label. The position of the tagged customer is indicated by the next two components; the index of the queue in which the tagged customer resides corresponds to the first component, the position of the tagged customer in the queue corresponds to the second. The queues are numbered as follows: 1 (CPU), 2 (Disk1), and 3 (Disk2). The state 00 is used to indicate that the tagged customer has departed from the system, i.e., has reached the absorbing state of the CTMC. States $(10000), (01000)$, and (00100) are the three absorbing states in the Markov chain, and are explicitly identified in the figure by the squares enclosed within the circles.

The three states explicitly identified in this figure by double circles are those states in which the tagged customer may arrive into the queuing system. These are states $(20012), (11011)$, and (10111), which correspond to the other job being at the CPU, Disk1, and Disk2, respectively. We can compute the absorption time distribution of the CTMC assuming that we start with any of these arrival states as the initial state. Each absorption time distribution represents the conditional response time distribution of the tagged customer, conditioned on a specific position of the other customer at the instant of arrival of the tagged customer into the queuing system.

Let I be the set of all states in the CTMC whose solution yields the response time distribution. The set of absorbing states in the CTMC is represented by A $(\subseteq I)$, and the set of states in which the tagged customer will find the network at the instant of arrival is represented by S $(\subseteq I)$. S is the set of all possible states of the network with one less customer for a closed queuing network. The random variable representing the response time for an arbitrary customer arriving when the queuing network is in state i, where $i \in S$, is R_i. Define $p_{ij}(t)$ as the transient probability of being in state j at time t given that the initial state of the CTMC in Fig. 14.2 is i. Then



State label : $(i\ j\ k\ l\ m)$
i: No. of jobs in CPU
j: No. of jobs in Disk1
k: No. of jobs in Disk2
l: Queue in which tagged customer is present (1=CPU, 2=Disk1, 3=Disk2)
m: Position of tagged customer in queue
$l = 0, m = 0$: Job has exited from the system

Fig. 14.2: CTMC model of the CSM for computing the response time distribution

$$P(R_i \le t) = \sum_{j \in A} p_{ij}(t).$$

To compute the unconditional response time distribution, we require the (steady-state) probabilities $\pi_i(n-1)$ that the tagged customer will see the network with the other $n-1$ customers in state i ($i \in S$) at the instant of arrival. These probabilities are computed based on a further CTMC.

In Fig. 14.3, this CTMC is shown for the case of $n = 2$ customers. It has three states corresponding to the non-tagged customer being present at the CPU, Disk1, and Disk2, respectively.

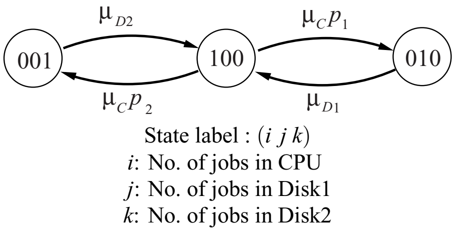

μ_{D2}
$\mu_C p_1$
001
100
010
$\mu_C p_2$
μ_{D1}
State label : $(i\ j\ k)$
i: No. of jobs in CPU
j: No. of jobs in Disk1
k: No. of jobs in Disk2

Fig. 14.3: CTMC model for computing the steady-state probabilities of the non-tagged customer

Let the random variable representing the unconditional response time be R. Then a general expression for the unconditional response time distribution is

$$\begin{aligned} P(R \le t) &= \sum_{i \in S} \pi_i(n-1) P(R_i \le t) \;=\; \sum_{i \in S} \pi_i(n-1) \sum_{j \in A} p_{ij}(t) \\ &= \sum_{j \in A} \sum_{i \in S} \pi_i(n-1) p_{ij}(t) \;=\; \sum_{j \in A} p_j(t), \end{aligned}$$

where $p_j(t)$ represents the unconditional transient probability of being in state j. For $n = 2$ customers, these probabilities are obtained by solving the CTMC in Fig. 14.2 for its transient probability vector at time t, given the initial probability of the state i $(\forall i \in S)$ of the CTMC is $\pi_i(n-1)$ and the initial probabilities of all the other states $(i \in I - S)$ are zero. The unconditional response time distribution is thus directly computed by assigning the initial probabilities for the CTMC and carrying out the transient analysis only once.

For this example, we set $\mu_C = 50.0, \mu_{D1} = 30.0, \mu_{D2} = 20.0, p_1 = 0.45$, and $p_2 = 0.3$. Figure 14.4 shows the response time distributions of the CSM for different numbers of customers n $(5, 10$ and $15)$. Note that when there are fewer customers competing for resources (i.e., the number of customers is smaller), a customer has a higher probability of completing by a given time t.

For a general method of constructing the two CTMCs for computing the response time distribution see [33], and for automated construction and solution of such models using stochastic Petri nets see [32]. See [7] and [12] for methods for efficiently analyzing response time densities in very large Markov and semi-Markov models.

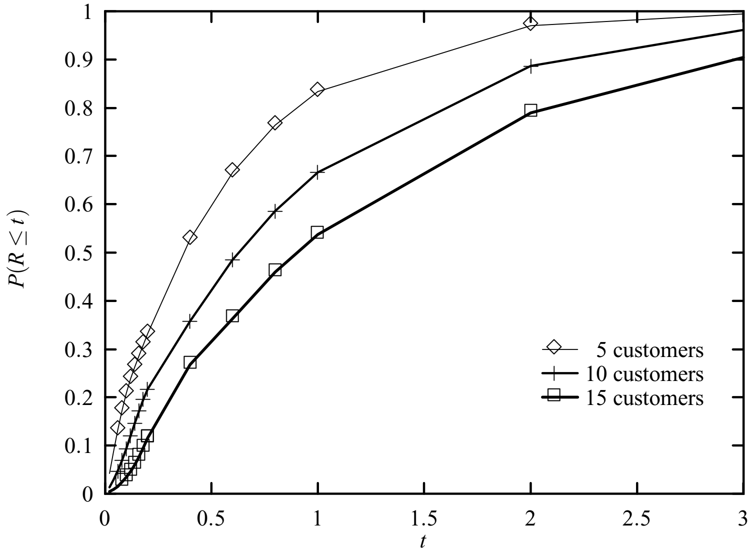


Fig. 14.4: Response time distribution of the CSM for different number of customers

14.3 Open Markovian networks of $M/M/c/b \leq \infty$ queues

As we have seen in the previous section, even for a rather simple closed system in which only few jobs are present, the total number of states that the system can take may be high. For open Markovian networks, in which the number of jobs in the system is not fixed, computing the response time distribution via the tagged customer approach basically involves an infinite number of states. However, there are methods of approximating the distribution that can give good results with much less effort. One such method uses the original network and knowledge of the response time distribution of its nodes to construct a CTMC for which the distribution of the time to reach the absorbing state can be solved to find the approximate response time distribution of the queuing network. In its simplest form, this method requires that the response times at the individual queues can be represented by the time until absorption in a CTMC. Fortunately, this is the case for many types of queues.

For five fairly simple types of queues, we present the CTMCs corresponding to their response time distributions in Sect. 14.3.1. These "response time blocks" can then be used to build the CTMC for the entire queuing network via the procedure laid out in Sect. 14.3.2. The approach is illustrated with three examples in Sect. 14.3.3.

14.3.1 Response time blocks

Throughout this section, we assume that all nodes use FCFS scheduling. Let λ and μ denote the arrival rate to the node and the service rate of each server of the node, respectively. To maintain stability, we assume that $\lambda < c\mu$, where c is the number of servers in the node. This is equivalent to the condition that the traffic intensity at the node, $\rho = \frac{\lambda}{c\mu}$, is smaller than one.

14.3.1.1 $M/M/1$

For an $M/M/1$, FCFS server with arrival rate λ and service rate μ, the response time distribution is given by Gross et al. [13] as

$$F(t) = 1 - \exp(-(\mu - \lambda)t),$$

assuming that $\lambda < \mu$. The response time block is shown in Fig. 14.5. State "In" indicates the starting state of the piece of the CTMC model representing the $M/M/1$ queue in the corresponding network model. The "Out" state will be either an "In" state for another network model node or the absorbing state representing the job leaving the network model. If the output of the queuing node can proceed on more than one path, then the "Out" state shown would actually be multiple states and the incoming arcs to these states would be weighted with the appropriate probabilities.

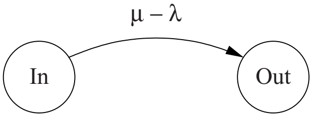


Fig. 14.5: The $M/M/1$ response time block

14.3.1.2 $M/M/\infty$

The $M/M/\infty$ server is the simplest node for which to find the response time distribution. Since there are always enough servers for any customer, the response time distribution is the same as the service time distribution,

$$F(t) = 1 - \exp(-\mu t).$$

Figure 14.6 shows the response time block for this distribution. It is very similar to Fig. 14.5, except that for the $M/M/\infty$ case the rate of leaving the "In" node is simply μ.

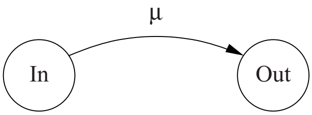


Fig. 14.6: The $M/M/\infty$ response time block

14.3.1.3 $M/M/c$

If there are assumed to be some finite number, c, of servers each having the same service rate, μ, and an infinite queue, we have an $M/M/c$, FCFS queue. The $M/M/c$, FCFS response time distribution given by Gross et al. [13] as

$$\begin{aligned} F(t) &= W_c(1-\exp(-\mu t)) \qquad (14.1)\\ &+ (1-W_c)\left[\frac{c\mu-\lambda}{(c-1)\mu-\lambda}[1-\exp(-\mu t)] - \frac{\mu}{(c-1)\mu-\lambda}[1-\exp(-(c\mu-\lambda)t)]\right], \end{aligned}$$

where

$$W_c = 1-\left[\frac{(c\rho)^c}{c!}\cdot\frac{1}{1-\rho}\right]\cdot\left[\sum_{k=0}^{c-1}\frac{(c\rho)^k}{k!}+\frac{(c\rho)^c}{c!}\cdot\frac{1}{1-\rho}\right]^{-1}$$

is the steady-state probability that a newly arriving job finds less than c jobs present in the node and therefore does not have to queue for service.

The distribution in Eq. (14.1) is a mixture of W_c fraction following an exponential distribution with parameter μ and $(1-W_c)$ fraction following a two-stage hypoexponential distribution with parameters μ and $c\mu-\lambda$. It can be described by the building block shown in Fig. 14.7. The upper path represents the exponentially distributed portion and the lower path is the hypoexponentially distributed portion. State T is strictly a transient state that is required to obtain the hypoexponential distribution.

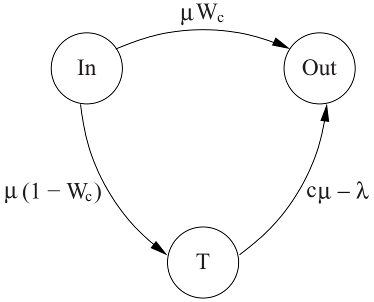


Fig. 14.7: The $M/M/c$, FCFS response time block

14.3.1.4 $M/M/c/b$

If we limit the previous case to a finite queue length (or buffer size) of b (where the queue includes the customer being serviced), we have the $M/M/c/b$, FCFS queue. The distribution given here is for an open network, assuming the job is accepted for service. In [43], the distribution for this case has been given as

$$\begin{aligned}F(t) &= \sum_{n=0}^{c-1} q_n(1-\exp(-\mu t))\\ &\quad + \sum_{n=c}^{b-1} q_n \left\{ \left(\frac{c}{c-1}\right)^{n-c+1} (1-\exp(-\mu t)) \right.\\ &\qquad \left. - \sum_{i=0}^{n-c} \left(\frac{c}{c-1}\right)^{n-c-i+1} \frac{1}{c} \left[1 - \sum_{j=0}^{i} \frac{(\mu ct)^j}{j!} \exp(-\mu ct)\right]\right\},\end{aligned}$$

where

$$q_n = \begin{cases} \dfrac{(c\rho)^n}{n!} \cdot \left[\displaystyle\sum_{k=0}^{c-1} \frac{(c\rho)^k}{k!} + \frac{c^c}{c!} \sum_{k=c}^{b-1} \rho^k\right]^{-1} & \text{if } n = 0, 1, ..., c-1, \\ \dfrac{c^c \rho^n}{c!} \cdot \left[\displaystyle\sum_{k=0}^{c-1} \frac{(c\rho)^k}{k!} + \frac{c^c}{c!} \sum_{k=c}^{b-1} \rho^k\right]^{-1} & \text{if } n = c, c+1, ..., b-1 \end{cases} \tag{14.2}$$

represents the conditional steady-state probability that a newly arriving job which is *not* blocked by the node due to a full buffer finds n other jobs present at the node.

This distribution is a mixture of an exponential distribution with parameter μ and $b-c$ hypoexponential distributions. Each hypoexponential distribution has a different number of phases. More specifically, the ith hypoexponential distribution ($i = 1, \ldots, b-c$) consists of $i+1$ phases; one phase has parameter μ, and the remaining i phases have parameter $c\mu$. Figure 14.8 shows the response time block representation of the distribution. Similar to the $M/M/c$ block, the states $T_2, \cdots, T_{b-c+1}$ are strictly transient states required to obtain the hypoexponential distributions. The probabilities V_j that appear in the transition rates are given by

$$V_j = \begin{cases} \displaystyle\sum_{n=0}^{c-1} q_n & \text{if } j = 1, \\ \dfrac{q_{c+j-2}}{\sum_{n=c+j-2}^{b-1} q_n} & \text{if } j = 2, ..., b-c+1. \end{cases}$$

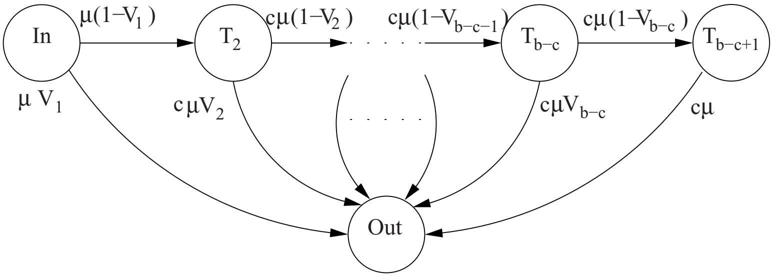


Fig. 14.8: The $M/M/c/b$ response time block

14.3.1.5 $M/M/1/b$

By limiting the $M/M/c/b$ case to a single server, we have an $M/M/1/b$, FCFS queue. The distribution given in this section has the same assumptions and qualifications as that in the $M/M/c/b$ section. The response time distribution for the $M/M/1/b$, FCFS case is given by Gross et al. [13] as

$$F(t) = \frac{1-\rho}{1-\rho^b}\sum_{n=0}^{b-1}\rho^n\left(1-\sum_{k=0}^{n}\frac{(\mu t)^k\exp(-\mu t)}{k!}\right).$$

Note that this distribution is a mixture of Erlang distributions where each has parameter μ and the number of phases varies from 1 to b. Setting $c = 1$ and

$$V_j = \frac{1-\rho}{1-\rho^{b-j+1}} \text{ if } 1 \le j \le b,$$

we can again use Fig. 14.8 to represent the $M/M/1/b$ response time block; see [48].

14.3.2 *Building the Markov chain from a queuing network*

With the building blocks for five types of queues described in Sect. 14.3.1, we can now outline a procedure for automatically mapping a queuing network to a response time CTMC.

Consider a network consisting of m queues $1,2,\ldots,m$, where each queue is of one of the types described in Sect. 14.3.1. Let λ_{0i} be the external arrival rate to node i in the queuing network. Suppose r_{i0} is the probability that a customer exits the network after receiving service at node i. Let $r_{ii'}$ be the probability that a customer proceeds to node i' after receiving service at node i. If queue i' is a finite capacity queue and it is full to its capacity, then we consider the following possibilities [36]: (1) The customer waits at queue i and is retried. (We call this policy WAIT.) (2) The customer is lost. (This policy is termed LOSS.) Suppose the probability of loss at queue i' is $p_{bi'}$. (Thus $p_{bi'} = 0$ for infinite capacity queues.) Then the procedure for building the overall CTMC is described as follows:

Step 1: Calculate effective arrival rates to each node of the queuing network. The effective arrival rate $\lambda_{i'}$ to node i' is given by [15]

$$\lambda_{i'} = \lambda_{0i'} + \sum_{i=1}^{m} \lambda_i (1 - p_{bi}) r_{ii'} \quad \text{for } i' = 1,2,...,m. \tag{14.3}$$

If all queues in the network are either infinite capacity queues or finite capacity queues with WAIT policy, i.e., if $p_{bi} = 0\ \forall i$, then (14.3) is a simple linear system of equations. However, things are more complicated if customers can get lost in the system. Let queue i of the network be an $M/M/c/b$ queue with LOSS policy and the same service rate μ_i for each of the c servers. Assuming that the arrivals to this queue are Poisson, the expression for p_{bi} is

$$p_{bi} = \frac{c^c \rho_i^b}{c!} \cdot \left[\sum_{k=0}^{c-1} \frac{(c\rho_i)^k}{k!} + \frac{c^c}{c!} \sum_{k=c}^{b} \rho_i^k \right]^{-1} .$$

Since $\rho_i = \frac{\lambda_i}{c\mu_i}$, substituting this expression into Eq. (14.3) results in a system of equations that may need to be solved using fixed-point iteration to yield the effective arrival rates. The buffer-full probabilities can then be calculated based on these values of the arrival rates. The effective arrival rates as well as the buffer-full probabilities may be approximate because we assume that the arrivals to all queues are Poisson, which is not necessarily the case.

Step 2: Create a CTMC with a state S_f. S_f is the state which denotes exit out of the queuing network. If the full buffer policy at any queue is LOSS then a state S_l is also created which denotes loss due to encountering full buffers.

For each queue i, create a set of states $S_i = \{S_{i1}, S_{i2}, \ldots, S_{in_i}\}$, consisting of all the states but the "Out" state contained in the response time block representation of this queue (see Sect. 14.3.1). Therefore,

$$n_i = \begin{cases} 1 & \text{if queue } i \text{ is } M/M/\infty \text{ or } M/M/1, \\ 2 & \text{if queue } i \text{ is } M/M/c, \\ b & \text{if queue } i \text{ is } M/M/1/b, \\ b-c+1 & \text{if queue } i \text{ is } M/M/c/b. \end{cases}$$

Then the state space of the CTMC is given by $S = (\bigcup_{i=1}^{m} S_i) \bigcup \{S_f, S_l\}$. Let $d = \max\{n_i \mid i = 1,2,\ldots,m\}$. Since

$$m < |S| \leq dm + 2,$$

the total number of states $|S|$ is linear in m.

Step 3: The response time distribution of the types of queues considered above are phase-type [42] and hence, in general, can be represented as follows: Let the sojourn time in the response time block state j of queue i follow an exponential distribution with rate μ_{ij} and let the probability of exiting the queue from that state be V_{ij}. Assume we define probability p_{bi} for all queues, i.e., it is zero for infinite capacity queues. Then the generator matrix $\mathbf{Q}$ of the CTMC defined in Step 2 is constructed based on the following principles.

The progress within the response time block of queue i, from state j to $j+1$, is preserved as in the block:

$$Q_{S_{ij},S_{i(j+1)}} = \mu_{ij}(1 - V_{ij}).$$

To represent routing from node i to i',

$$Q_{S_{ij},S_{i'1}} = \mu_{ij} V_{ij} r_{ii'} (1 - p_{bi'}).$$

For denoting exit from the network,

$$Q_{S_{ij},S_f} = \mu_{ij} V_{ij} r_{i0}.$$

If there is any queue with LOSS policy in the queuing network, then we also have, in addition to the above,

$$Q_{S_{ij},S_l} = \mu_{ij} V_{ij} \sum_{i' \neq i} r_{ii'} p_{bi'}.$$

Note that in case of WAIT, we assume that the time until retry is the same as the time to exit queue i starting from the state from which exit was attempted. We also assume that on retry the queue to which the job is routed is sampled again.

The rest of the entries of $\mathbf{Q}$ are zeros except for the diagonal entries, which are

$$Q_{S_i,S_i} = - \sum_{S_{i'} \in S, S_{i'} \neq S_i} Q_{S_i,S_{i'}}.$$

Step 4: If we define $\lambda = \sum_{i=1}^{m} \lambda_{0i}$, then λ_{0i}/λ is the probability that an external customer arrives at queue i. Then the initial state probability is set to λ_{0i}/λ for state S_{i1} and to 0 for all other states.

The cumulative distribution function of the response time R of the network can now be found by computing $P_{S_f}(t)$, the probability of being in state S_f at time t. Since S_f is an absorbing state, this probability is equal to the probability of exiting the network before time t, given that the customer entered it at time 0. Note that if we have LOSS, this distribution will be *defective*, i.e., $\lim_{t\to\infty} P(R \leq t)$ $= \lim_{t\to\infty} P_{S_f}(t) < 1$. This is because there is a non-zero probability that the customer will never reach the state S_f.

Let us assume

- that successive service times are independent (Kleinrock's independence assumption),
- that the independent flow time approximation as described in Sect. 14.1 holds,

- that external arrivals are Poisson,
- that all queues are single server with exponentially distributed service time and FCFS queuing discipline,
- that all queues are infinite capacity queues,
- that the network is feed-forward, and
- that all the paths in the network are overtake-free [6].

If the above conditions are met, arrivals at all queues are Poisson, and further, successive sojourn times in a path are independent. In that case, $P_{S_f}(t)$ as computed by the above algorithm gives the exact distribution of sojourn time. When any of the above conditions are violated, the absorption time distribution is an approximation. The successive sojourn times in queues in a path with overtaking or feedback are correlated [6], thus violating the implicit independence assumption in our method.

14.3.3 Examples

We will now illustrate the approach developed above using a computer system example and a distributed system example. We also show how this methodology can be used to find the distribution of the sample average of response times.

14.3.3.1 Computer system

Consider the simple model of a computer system (Fig. 14.9), which is comprised of two $M/M/1$ queues.

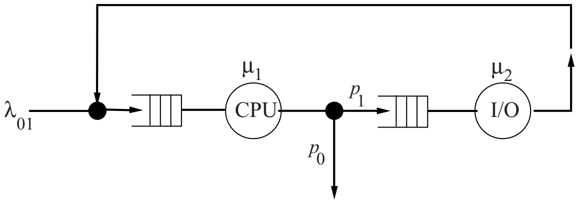


Fig. 14.9: Model I, an open network with feedback

Defining the effective arrival rates to the CPU queue and the I/O queue to be λ_1 and λ_2, respectively, we follow Step 1 and formulate the equation system

$$\lambda_1 = \lambda_{01} + \lambda_2,$$
$$\lambda_2 = \lambda_1 p_1.$$

This system of equations can easily be solved for the effective arrival rates,

$$\lambda_1 = \frac{\lambda_{01}}{1-p_1} = \frac{\lambda_{01}}{p_0} \quad \text{and}$$
$$\lambda_2 = \frac{p_1\lambda_{01}}{p_0}.$$

We now build the CTMC from the queuing network according to Steps 2 and 3. The state space S consists of the "In" nodes of the two $M/M/1$ response time blocks, denoted by S_{11} and S_{21}, respectively, as well as the exit node S_f. Since both queues have infinite capacity, $p_{b1} = p_{b2} = 0$. The rate of leaving the "In" state of an $M/M/1$ response time block is calculated as the service rate minus the arrival rate; consequently, we have $\mu_{i1} = (\mu_i - \lambda_i)$ for $i = 1,2$. Moreover, the probabilities of leaving the current node is equal to one for the states S_{11} and S_{21}, i.e., $V_{11} = V_{21} = 1$. According to the model, a job leaving the CPU can either exit the network with probability $r_{10} = p_0$ or progress to the I/O queue with probability $r_{12} = p_1$; all jobs finished at the I/O queue return to the CPU queue; i.e., $r_{21} = 1$ and $r_{20} = 0$. Therefore, the only off-diagonal entries of $\mathbf{Q}$ that are not equal to zero are

$$\begin{aligned} Q_{S_{11},S_{21}} &= (\mu_1 - \lambda_1)p_1, \\ Q_{S_{11},S_f} &= (\mu_1 - \lambda_1)p_0, \\ Q_{S_{21},S_{11}} &= \mu_2 - \lambda_2. \end{aligned}$$

The corresponding CTMC is depicted in Fig. 14.10.

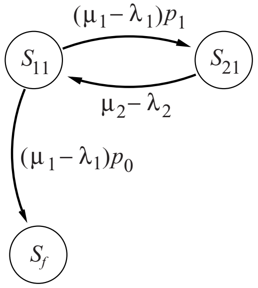


Fig. 14.10: CTMC corresponding to the response time distribution of model I

It can be shown that the expected response time in this queuing model is given by [42, p. 561]

$$E(R) = \frac{1}{p_0\mu_1 - \lambda_{01}} + \frac{1}{\frac{p_0\mu_2}{p_1} - \lambda_{01}}.$$

Obviously, the same expected response time is obtained in the simple network without feedback shown in Fig. 14.11. However, as we will see, the cumulative distribution function of the response time in this "equivalent" model II is different from the one in the original model I.

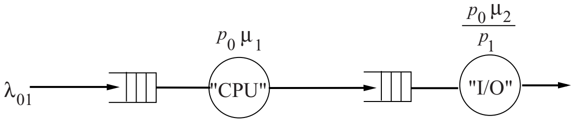


Fig. 14.11: Model II, an “equivalent” network without feedback

The response time CTMC for queuing network model II is built from two $M/M/1$ response time blocks with $\mu_{11} = p_0\mu_1 - \lambda_{01}$ and $\mu_{21} = \frac{p_0\mu_2}{p_1} - \lambda_{01}$. As before, $p_{bi} = 0$ and $V_{i1} = 1$ for $i = 1,2$. Since all jobs proceed from queue 1 to queue 2 and then leave the network, $r_{12} = r_{20} = 1$. The non-zero off-diagonal elements of **Q** are thus

$$\begin{aligned} Q_{S_{11},S_{21}} &= p_0\mu_1 - \lambda_{01}, \\ Q_{S_{21},S_f} &= \frac{p_0\mu_2}{p_1} - \lambda_{01}. \end{aligned}$$

The CTMC is shown in Fig. 14.12.

For a specific example, we assume that the arrival rate from outside the system is $\lambda_{01} = 1$ job per second and that a job leaves the system with a probability of $p_0 = 0.2$ after being processed at the CPU. We further assume that the CPU can process jobs at a rate of $\mu_1 = 10$ jobs per second and the I/O can process jobs at a rate of $\mu_2 = 5$ jobs per second.

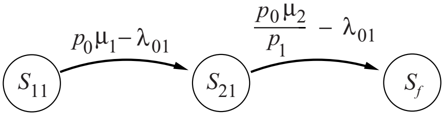

$p_0\mu_1 - \lambda_{01}$
$\frac{p_0\mu_2}{p_1} - \lambda_{01}$
S_{11}
S_{21}
S_f

Fig. 14.12: CTMC corresponding to the response time distribution of model II

The CTMCs of model I and model II were solved using the SHARPE [37] tool to obtain the distributions of time to absorption in state S_f. For comparison, the queuing network models were also implemented and the response time distribution found using the simulation tool HyPerformix Workbench[1].

The CTMC solutions as well as the simulations found the expected response time to be 5 seconds for both network models. All response time distributions obtained are plotted in Fig. 14.13. Note that for the model I (Fig. 14.9) the CTMC approximation shows the distribution to be slightly lower than the resultant distribution of the simulation for values of $t < 12$, and slightly higher for the remaining values of t. For model II (Fig. 14.11), the results from the CTMC approximation and the simulation model are in agreement for all t. This can be explained by the fact that the CTMC

[1] Registered trademark of HyPerformix, Inc.

approach gives exact results when the queuing network is an open Jacksonian feed-forward network with overtake-free conditions [29]. However, comparing the results from the “equivalent” network to the original, we see quite a bit of difference in the response time distributions.

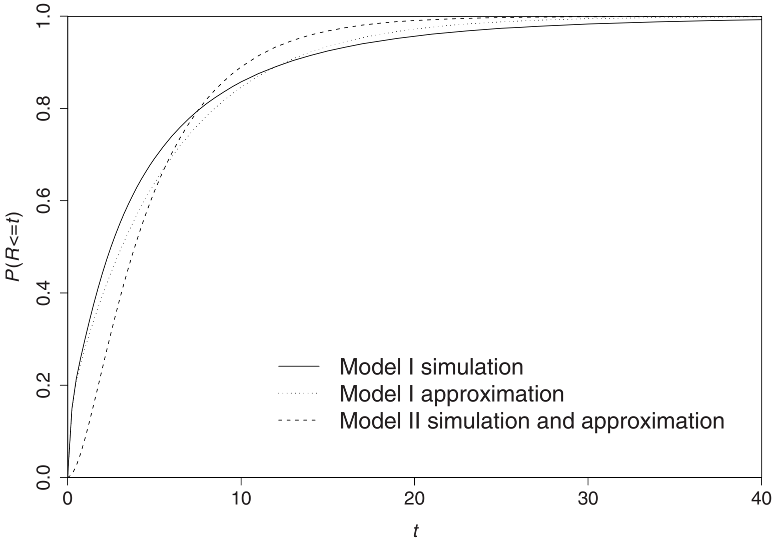


Fig. 14.13: Response time distribution for the networks of Figs. 14.9 and 14.11

14.3.3.2 Distributed system

Consider a distributed system (Fig. 14.14) in which users send requests from terminals (T) at the rate λ. A job first obtains service from a front-end server (F) and may exit the system with probability p_0 after completion of service. With probability p_1, it proceeds to the communications server (C). After completion of service it may go back to the front-end server with probability p_2, or proceed to a database server (D) with probability p_3 or to a general-purpose server (P) with probability p_4.
The terminals (T) are assumed to be $M/M/\infty$ servers having service rate μ_T. F is an $M/M/c_F$ server with each of the c_F servers having service rate μ_F. C is assumed to be a single server ($M/M/1$) having service rate μ_C. D is assumed to be a single server device having a finite capacity ($M/M/1/b_D$). The service rate for the database server is μ_D. P is assumed to be a multi-server having a finite capacity ($M/M/c_P/b_P$). The service rate of each server is assumed to be μ_P. We shall evaluate this system first assuming the WAIT policy at both finite capacity queues and then assuming the LOSS policy at both finite capacity queues.

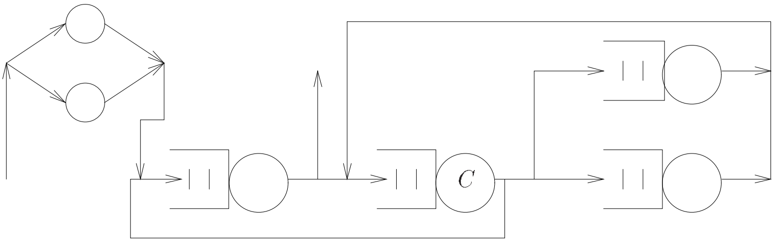


Fig. 14.14: Distributed system queuing network

WAIT

We first describe the results under the WAIT policy. For this network, WAIT policy means that jobs that cannot go to the D or P queues due to lack of buffer availability will remain at the communications server to be retried. With this assumption, on solving Eq. (14.3), we obtain the following values for effective arrival rates λ_i to each of the queues $i = F, C, D, P$ [48]:

$$\lambda_F = \frac{\lambda}{p_0}, \qquad \lambda_C = \frac{p_1}{p_0 p_2} \cdot \lambda,$$
$$\lambda_D = \frac{p_1 p_3}{p_0 p_2} \cdot \lambda, \qquad \lambda_P = \frac{p_1 p_4}{p_0 p_2} \cdot \lambda.$$

Figure 14.15 shows the CTMC corresponding to the response time distribution of this queuing network when the capacity at each of the queues D and P is 4, and $c_P = 2$.

State T, matching the "In" state of the $M/M/\infty$ response time block (Fig. 14.6), is the starting state of this CTMC. Likewise, states $F1$ and $F2$ correspond to the "In" and T states, respectively, of the building block shown in Fig. 14.7, while states $D1$, $D2$, $D3$ and $D4$ are related to the "In" and T_j states for $j = 2, 3, 4$ of the $M/M/c/b$ response time block of Fig. 14.8. The other states can be similarly identified. State "Done" corresponds to the state S_f described in Sect. 14.3.2. The probabilities p_{bD} and p_{bP} are the buffer full probabilities of queues D and P, respectively. Finally, ρ_D denotes the utilization in the database server node, while the variables q_n ($n = 0, 1, 2, 3$) denote the conditional probabilities that a job arriving and not being blocked at the general-purpose server finds n other jobs at this node, as defined in Eq. (14.2).

For $\mu_T = 0.2$, $\mu_F = 1.5/4$, $\mu_C = 1$, $\mu_D = 0.2$, $\mu_P = 0.05$, $c_F = 4$, $c_P = 2$, $p_0 = 0.5$, $p_1 = 0.5$, $p_2 = 0.46$, $p_3 = 0.33$ and $p_4 = 0.21$, the distribution of the response time in this network is shown in Fig. 14.16(a). The figure compares these results with the simulation using RESQ, the queuing network modeling environment developed at IBM [27]. The simulation was done using the regenerative method and provided 99% confidence intervals with maximum width 0.008.

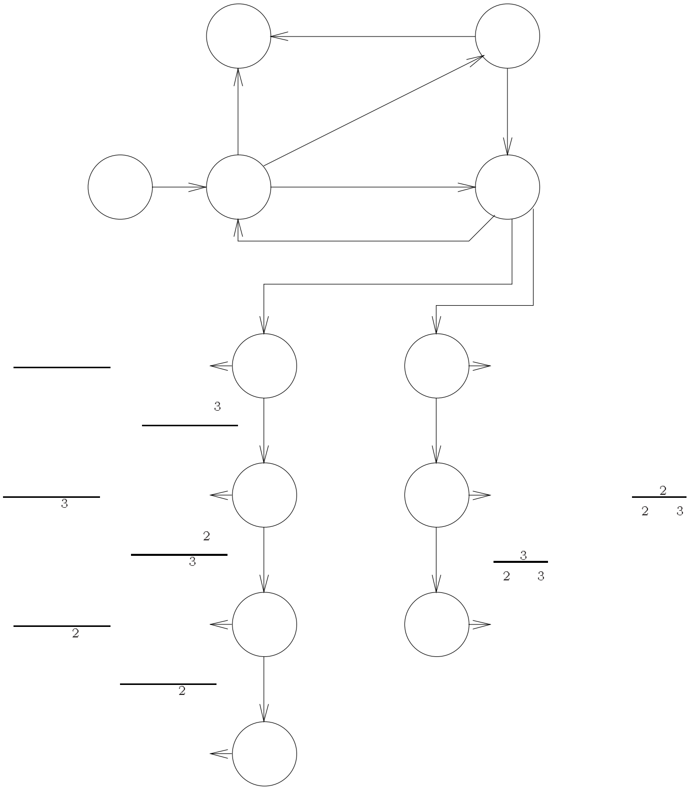


Fig. 14.15: CTMC corresponding to response time distribution for WAIT

(Since the confidence intervals obtained are very small, we have used the midpoint of the intervals to plot the response time distribution.)

Figure 14.16(b) compares the response time distribution of this system for various arrival rates.

LOSS

Under the LOSS policy we have to take into account the buffer-full probabilities p_{bD} and p_{bP} when setting up equation system (14.3). Therefore, the effective arrival rates are different from the ones in the WAIT policy case:

$$\lambda_F = \frac{p_0p_2 + p_1p_2 + p_{bD}p_3 + p_4p_{bP}}{p_0p_2 + p_3p_{bD} + p_4p_{bP}} \cdot \lambda, \qquad \lambda_C = \frac{p_1}{p_0p_2 + p_3p_{bD} + p_4p_{bP}} \cdot \lambda,$$

$$\lambda_D = \frac{p_1p_3}{p_0p_2 + p_3p_{bD} + p_4p_{bP}} \cdot \lambda, \qquad \lambda_P = \frac{p_1p_4}{p_0p_2 + p_3p_{bD} + p_4p_{bP}} \cdot \lambda.$$

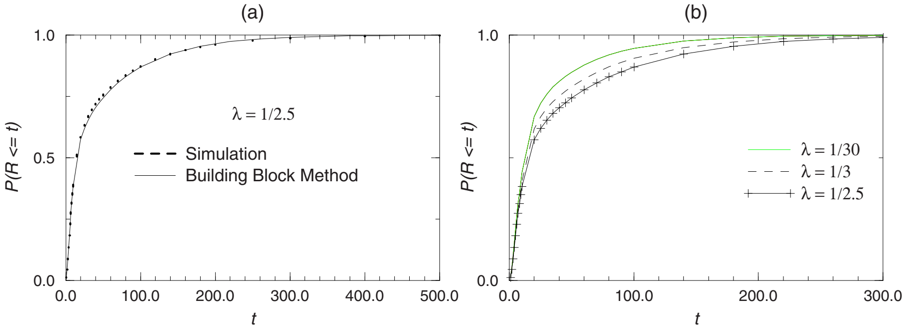


Fig. 14.16: (a) Response time distribution of distributed system: comparison with simulation (b) Response time distribution for various values of λ

Since jobs leaving node C can get lost at node D or node P, the CTMC constructed for the WAIT policy has to be extended by a state "Lost", corresponding to the state S_l described in Sect. 14.3.2, and a transition from state $C1$ to this "Lost" state. The resulting CTMC corresponding to the response time distribution under the LOSS policy is shown in Fig. 14.17.

Figure 14.18(a) shows a comparison of the response time distribution obtained by solving the Markov model with the one obtained by RESQ simulation. In the LOSS case deriving confidence intervals using RESQ would have taken prohibitively long time, hence the plot shows point estimates. Notice that the distribution in this case is defective, and thus $\lim_{t\to\infty} P(R \le t) < 1$.

Figure 14.18(b) shows how the response time distribution improves with increase in buffer size, since fewer jobs get lost. Note that the conditional response time, given that the job does not get lost, will degrade with larger buffer space, since the length of the queue will increase.

14.3.3.3 Distribution of the response time sample mean

In some applications, we are not only interested in the distribution of the response times, but in the distribution of the sample mean calculated from n observed response times. For example, in [3] we examined a multi-tier e-commerce application consisting of 16 CPUs; the normal system behavior could be represented by an $M/M/c$ queue with $c = 16$. However, due to garbage collection events and kernel overhead, the system sometimes showed severe performance degradation. The only remedy in such a situation was to "rejuvenate" the system by terminating all threads in execution, which freed the resources held by these threads. In [3], we developed several algorithms triggering rejuvenation based on recent observations of response times. Since the rejuvenation of the system incurs costs (e.g., lost transactions), the challenge was to distinguish the sustained performance deterioration

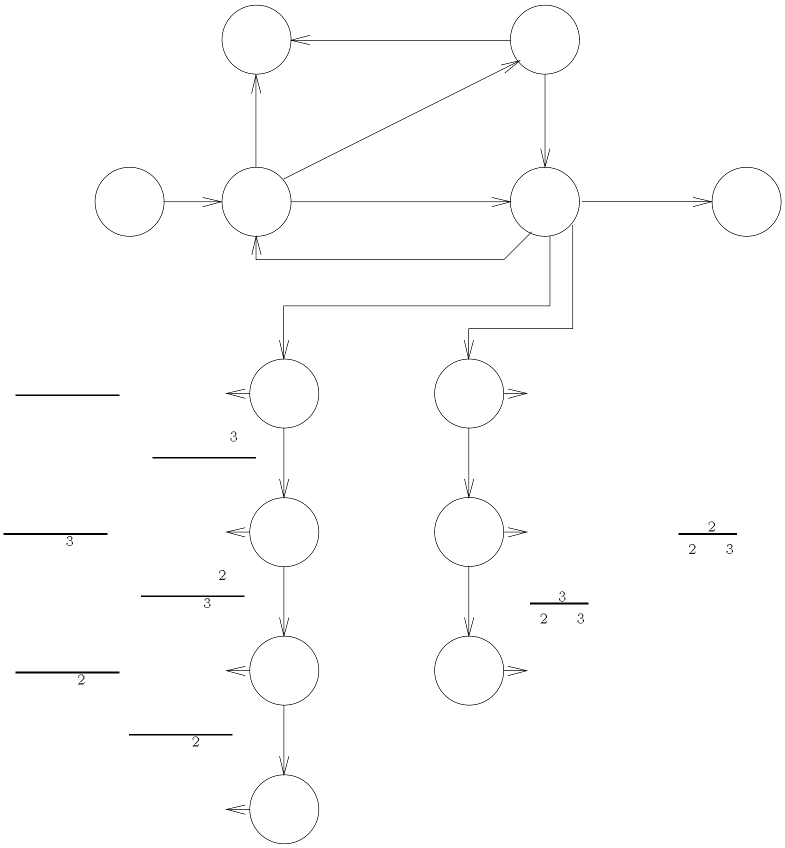


Fig. 14.17: CTMC corresponding to response time distribution for LOSS

from short-term increases in the observed response times. One of the algorithms studied employed the sample means of n subsequently observed response times in order to smooth out sporadically occurring large values of the response time.

Assuming that the observable response times R_i are independently and identically distributed (e.g., because the time lag between the collection of two response times via probing is sufficiently large), then the sample mean $\bar{R}_n = \frac{1}{n}\sum_{i=1}^{n} R_i = \sum_{i=1}^{n} \frac{R_i}{n}$ follows a phase-type distribution. Obviously, each individual summand $\frac{R_i}{n}$ is the response time of a job in an $M/M/c$, FCFS queue with arrival rate $n\lambda$ and service rate $n\mu$. Its distribution can then be represented by an adapted version of the $M/M/c$ building block (see Fig. 14.7), in which all transition rates are multiplied by n. Therefore, $\bar{R}_n$ corresponds to the time until reaching the absorbing state S_f in a simple network of n such $M/M/c$, FCFS queues, in which a job upon leaving queue $i < n$ proceeds to queue $i+1$. Note that, unlike in the previous examples, there is no physical queuing network corresponding to this network of queues.

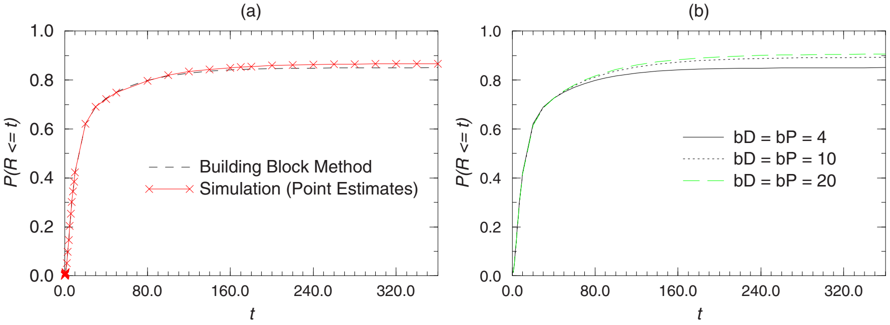


Fig. 14.18: (a) Response time distribution of distributed system: comparison with simulation (b) Response time distribution for various values of b_D,b_P

Due to the linear structure ($r_{i,i+1} = 1$ for $i = 1,...,n-1$) and the fact that no jobs are lost ($p_{bi} = 0$ for $i = 1,...,n$), the effective arrival rates are identical for all queues: $\lambda_i = n\lambda$. The sojourn time in states S_{i1} (corresponding to the "In" state in the response time block of queue i) and S_{i2} (corresponding to T state in the response time block of queue i) are exponential with parameters $\mu_{i1} = n\mu$ and $\mu_{i2} = n(c\mu - \lambda)$, respectively. The probability of leaving the queue from state S_{i1} is $V_{i1} = W_c$, while all jobs leave the queue from state S_{i2}, i.e., $V_{i2} = 1$. Upon leaving queue n, a job is routed to the absorbing state S_f, i.e., $r_{n0} = 1$. The CTMC corresponding to the distribution of the average response time thus consists of $2n+1$ states and has a generator matrix $\mathbf{Q}$ with non-zero off-diagonal elements

$$\begin{aligned}
Q_{S_{i1},S_{i+1,1}} &= n\mu W_c && \text{for } i = 1,...,n-1,\\
Q_{S_{i1},S_{i2}} &= n\mu(1-W_c) && \text{for } i = 1,...,n,\\
Q_{S_{i2},S_{i+1,1}} &= n(c\mu-\lambda) && \text{for } i = 1,...,n-1,\\
Q_{S_{n1},S_f} &= n\mu W_c,\\
Q_{S_{n2},S_f} &= n(c\mu-\lambda).
\end{aligned}$$

It is shown in Fig. 14.19.

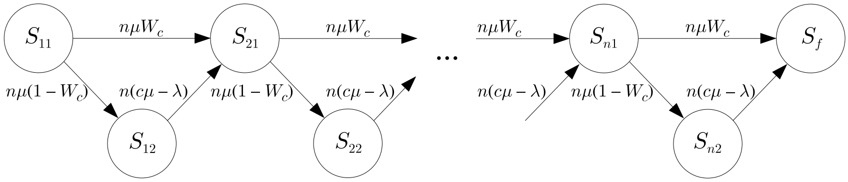


Fig. 14.19: CTMC corresponding to the distribution of the average response time $\bar{R}_n = \frac{1}{n}\sum_{i=1}^{n} R_i$

Solving this CTMC with the SHARPE [37] tool, we can obtain the cumulative distribution function of the sample mean, $F_{\bar{R}_n}(t) = P(\bar{R}_n \leq t)$. Due to the relationship

$$f_{\bar{R}_n}(t) = p_{n1}(t) \cdot n\mu W_c + p_{n2}(t) \cdot n(c\mu - \lambda),$$

with $p_{ij}(t)$ denoting the probability that the process is in state S_{ij} at time t, we can derive the probability density function $f_{\bar{R}_n}(t)$ by assigning the reward rates $n\mu W_c$ and $n(c\mu - \lambda)$ to the states S_{n1} and S_{n2}, respectively.

For different values of n, Fig. 14.20 shows this probability density function, based on a server with $c = 16$ CPUs, an arrival rate of $\lambda = 1.6$ jobs per second and a service rate of $\mu = 0.2$ jobs per second.

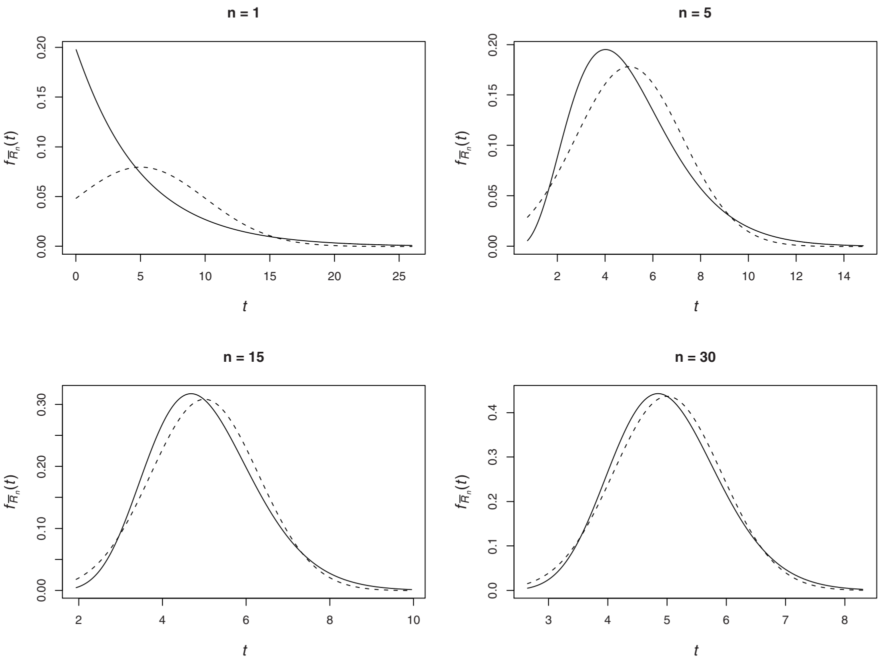


Fig. 14.20: Probability density function of average response time $\bar{R}_n$ for $n = 1, 5, 15, 30$ and corresponding approximating normal densities $f_N(t; \mu_{\bar{R}_n}, \sigma^2_{\bar{R}_n})$; $\lambda = 1.6, \mu = 0.2$

As a dashed line, each plot includes the probability density function of the corresponding normal distribution, with the same expected value and variance as the respective sample mean. The figures show how the sample mean converges against the normal distribution, as stated by the central limit theorem.

14.4 Open Markovian networks of queues with general PH service time distributions

The approach for deriving a CTMC for the response time distribution described in Sect. 14.3.2 can be employed if each of the nodes is of one of the five types presented in Sect. 14.3.1. However, there are additional types of queues for which the response time distribution can be represented by a CTMC. Important examples are $M/PH/1$ and $M/PH/\infty$ queues, in which the service time follows a phase-type (PH) distribution. This kind of distribution as well as the two types of queues will be discussed in the following section.

14.4.1 Building blocks with general PH service time distributions

We begin be reviewing the definition of the phase-type distribution [34]. Consider a CTMC on the states $\{1,\dots,n+1\}$ with infinitesimal generator

$$\mathbf{Q} = \begin{bmatrix} \mathbf{T} & \tau' \\ \mathbf{0} & 0 \end{bmatrix},$$

where the $n \times n$ matrix $\mathbf{T}$ satisfies $T_{ii} < 0$, for $1 \le i \le n$ and $T_{ij} \ge 0$ for $i \ne j$. Further, $\mathbf{T}\mathbf{e}' + \tau' = \mathbf{0}'$. Here and in the following, all vectors are by default row vectors, and column vectors are expressed as transposed vectors. For example, τ' is a column vector of length n, while $\mathbf{e}$ represents a row vector of n ones. The initial probability vector of the CTMC with infinitesimal generator matrix $\mathbf{Q}$ is given by $(\alpha, 1-\alpha\mathbf{e}')$, where α is a row vector of length n and $1-\alpha\mathbf{e}'$ represents the probability for the CTMC to start out in the absorbing state $n+1$. It is assumed that the states $1,\dots,n$ are all transient so that the process will always eventually reach the absorbing state. A necessary and sufficient condition for this is that the matrix $\mathbf{T}$ be non-singular. As defined by Neuts [34, p. 45], a probability distribution $F(.)$ on $[0,\infty)$ is a PH distribution if and only if it is the distribution of time until absorption in a finite CTMC of the type defined above. The pair $(\alpha,\mathbf{T})$ is called a representation of $F(.)$.

In the context of queuing networks, PH distributions are of relevance, because the queues $M/PH/\infty$ and $M/PH/1$, in which the arrival process is Poisson and the service time follows a phase-type distribution, have a phase-type response time distribution. This means that (like for the five types of queues discussed in Sect. 14.3.1) the response time distributions of $M/PH/\infty$ and $M/PH/1$ queues can be represented by CTMCs. We thus add these queues to our list of building blocks.

The case of an $M/PH/\infty$ queue is very simple: The Markov submodel for its response time is simply that one corresponding to the phase-type service time distribution.

The derivation of the Markov submodel is more complicated for an $M/PH/1$ queue. Suppose that the service time distribution of the queue has representation $(\alpha,\mathbf{T})$. According to Neuts [34, p. 57], the stationary waiting time distribution is then phase-type with representation

$$\beta = \rho\theta, \quad \mathbf{L} = \mathbf{T} + \rho\tau'\theta,$$

where $\theta = (\theta_1, \ldots, \theta_n)$ is the stationary probability vector of the CTMC with infinitesimal generator matrix $\mathbf{T} + \tau'\alpha$. This means that the stationary waiting time distribution is the time until absorption in the CTMC with infinitesimal generator matrix

$$\mathbf{Q}^* = \begin{bmatrix} \mathbf{L} & \delta' \\ \mathbf{0} & 0 \end{bmatrix}$$

and initial probability vector $(\beta, 1-\beta\mathbf{e}')$, where $\mathbf{L}\mathbf{e}' + \delta' = \mathbf{0}'$. From this representation a Markov submodel can be built.

14.4.2 Building the Markov chain

If nodes of the open queuing network are $M/PH/1$ or $M/PH/\infty$ type queues, we must slightly change the mapping procedure described in Sect. 14.3.2.

If queue i in the network is $M/PH/\infty$ with $(\alpha^{(i)}, \mathbf{T}^{(i)})$ denoting the phase-type representation of its service time, then create a set of states $\{S_{i1}, \ldots, S_{in_i}\}$, where n_i is the number of rows (or columns) of the matrix $\mathbf{T}^{(i)}$.

If queue i is an $M/PH/1$ queue, then create the states $\{W_{i1}, \ldots, W_{in_i}\}$ related to the PH waiting time distribution with representation $(\beta^{(i)}, \mathbf{L}^{(i)})$, and the states $\{S_{i1}, \ldots, S_{in_i}\}$ related to the PH service time distribution with representation $(\alpha^{(i)}, \mathbf{T}^{(i)})$. These states must be added to the total state space S in Step 2 of the mapping procedure (Sect. 14.3.2). In the following, we assume that $\alpha^{(i)}\mathbf{e}' = 1$; i.e., the service time of a queue cannot be zero.

When deriving the generator matrix $\mathbf{Q}$ in Step 3 of the procedure, rates for transitions involving $M/PH/1$ or $M/PH/\infty$ type queues are calculated as follows:

- Let queue i be an $M/PH/1$ queue. The transitions within the Markov submodel remain unchanged; i.e., each transition rate is given by the respective off-diagonal element in the matrix $\mathbf{L}^{(i)}$ or $\mathbf{T}^{(i)}$:

$$\begin{aligned} Q_{W_{ij}, W_{ij'}} &= L^{(i)}_{j,j'} \quad j \neq j', \\ Q_{S_{ij}, S_{ij'}} &= T^{(i)}_{j,j'} \quad j \neq j'. \end{aligned} \tag{14.1}$$

  To denote service after waiting,

$$Q_{W_{ij}, S_{ij'}} = \delta^{(i)}_j (\alpha^{(i)}_{j'})'.$$

  Now we consider the following cases for representing progress to the next queue i':

  – If queue i' is $M/M/c/b \leq \infty$,

$$Q_{S_{ij},S_{i'1}} = \tau_j^{(i)} r_{ii'}(1-p_{bi'}). \tag{14.2}$$

- If queue i' is $M/PH/\infty$,

$$Q_{S_{ij},S_{i'j'}} = \tau_j^{(i)} r_{ii'} \alpha_{j'}^{(i')}. \tag{14.3}$$

- If queue i' is also $M/PH/1$, then the job may either have to wait at this queue,

$$Q_{S_{ij},W_{i'j'}} = \tau_j^{(i)} r_{ii'} \beta_{j'}^{(i')}, \tag{14.4}$$

  or it may directly proceed to one of the states related to the service time distribution,

$$Q_{S_{ij},S_{i'j'}} = \tau_j^{(i)} r_{ii'}(1-\beta^{(i')}\mathbf{e}')\alpha_{j'}^{(i')}. \tag{14.5}$$

- If there are finite capacity queues with LOSS policy in the queuing network,

$$Q_{S_{ij},S_l} = \tau_j^{(i)} \sum_{i' \neq i} r_{ii'} p_{bi'}. \tag{14.6}$$

- If queue i is $M/PH/\infty$, Eq. (14.1) remains valid. For representing progress to the next queue, Eqs. (14.2)–(14.6) remain valid.
- If queue i is $M/M/c/b$, then there are two cases involving queues with PH service time distributions:
  - If queue i' is $M/PH/\infty$,

$$Q_{S_{ij},S_{i'j'}} = \mu_{ij} V_{ij} r_{ii'} \alpha_{j'}^{(i')}.$$

  - If queue i' is $M/PH/1$, then the job may proceed either to one of the states related to the waiting time distribution,

$$Q_{S_{ij},W_{i'j'}} = \mu_{ij} V_{ij} r_{ii'} \beta_{j'}^{(i')},$$

    or to one of the states related to the service time distribution,

$$Q_{S_{ij},S_{i'j'}} = \mu_{ij} V_{ij} r_{ii'}(1-\beta^{(i')}\mathbf{e}')\alpha_{j'}^{(i')}.$$

- The initial distribution of the CTMC also needs modification: If queue i is $M/PH/\infty$, then the probability of starting in state S_{ij} is $\lambda_{0i}\alpha_j^{(i)}/\lambda$. If queue i is $M/PH/1$, then the probability of starting in state W_{ij} is $\lambda_{0i}\beta_j^{(i)}/\lambda$ and the probability of starting in state S_{ij} is $\lambda_{0i}(1-\beta^{(i)}\mathbf{e}')\alpha_j^{(i)}$.

14.4.3 Example: CPU and disk queuing system

Consider a CPU and disk system as depicted in Fig. 14.21(a). Suppose the service time distribution at the CPU is hyperexponential with two stages (Fig. 14.21(b)). Suppose the service time at the disk is branching Erlang (Fig. 14.21(c)). Let the arrival process to the system be Poisson with rate λ. The response time distribution for this queuing network can be computed approximately with the CTMC approach.

First, the effective arrival rates to the CPU and disk are given by $\lambda_C = \lambda/(1-p_d)$ and $\lambda_D = p_d\lambda/(1-p_d)$. The phase-type representation $(\alpha_C, \mathbf{T}_C)$ for the hyperexponential distribution is $\alpha_C = (\alpha_{C1}, \alpha_{C2})$ and

$$\mathbf{T}_C = \begin{bmatrix} -\mu_1 & 0 \\ 0 & -\mu_2 \end{bmatrix}.$$

Using Neuts' theorem, we can derive the phase-type representation of the waiting time distribution of a customer at the CPU as

$$\beta_C = \left(\frac{\lambda_C\alpha_{C1}}{\mu_1}, \frac{\lambda_C\alpha_{C2}}{\mu_2}\right), \qquad \mathbf{L}_C = \begin{bmatrix} -\mu_1 + \lambda_C\alpha_{C1} & \dfrac{\lambda_C\mu_1\alpha_{C2}}{\mu_2} \\ \dfrac{\lambda_C\mu_2\alpha_{C1}}{\mu_1} & -\mu_2 + \lambda_C\alpha_{C2} \end{bmatrix}.$$

The phase-type representation of the service time at the disk is $\alpha_D = (1,0)$ and

$$\mathbf{T}_D = \begin{bmatrix} -\mu_3 & \mu_3(1-p) \\ 0 & -\mu_4 \end{bmatrix}.$$

The phase-type waiting time at the disk is given by

$$\beta_D = \left(\frac{\lambda_D}{\mu_3}, \frac{\lambda_D(1-p)}{\mu_4}\right), \qquad \mathbf{T}_D = \begin{bmatrix} -\mu_3 + \lambda_D p & \mu_3(1-p)\left(1 + \dfrac{\lambda_D p}{\mu_4}\right) \\ \dfrac{\lambda_D\mu_4}{\mu_3} & -\mu_4 + \lambda_D(1-p) \end{bmatrix}.$$

The CTMC corresponding to the response time distribution of the CPU and disk queuing system is depicted in Fig. 14.21(d). The response time distribution in this queuing network is approximated by computing the distribution of time to absorption, i.e., of reaching state S_F.

For the parameters $\lambda = 1$, $p_d = 0.7$, $\alpha_{C1} = 0.3$, $\alpha_{C2} = 0.7$, $\mu_{C1} = 6.67$, $\mu_{C2} = 10$ and for different disk service rates, Fig. 14.22 compares the response time distributions derived via the CTMC method with simulations. The vertical bars next to the solid lines denote 95% confidence intervals.

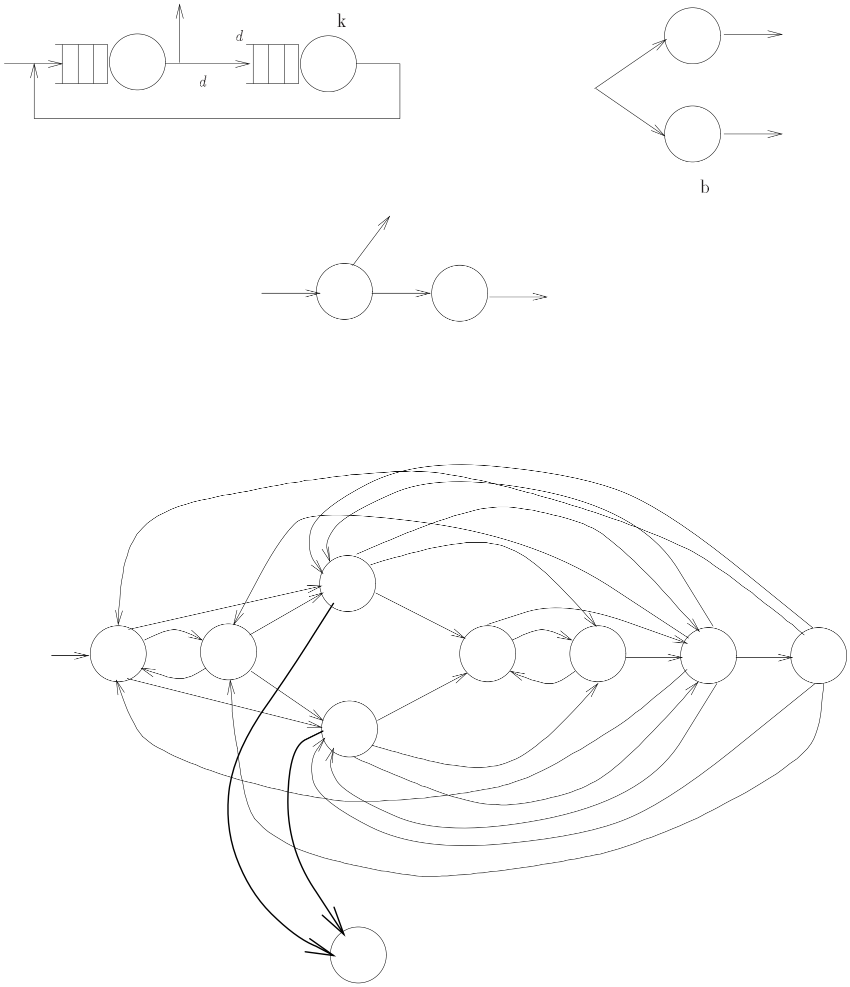


Fig. 14.21: (a) CPU-disk queuing system (b) Service time distribution at the CPU (c) Service time distribution at the disk (d) CTMC corresponding to the response time distribution

14.5 Non-Markovian networks

Approximating response time distributions for Markovian networks using the techniques previously described has been shown to give accurate results with much less computational effort than required for exact analysis. In this section, we apply similar paradigms in computing approximations to the response time distribution of

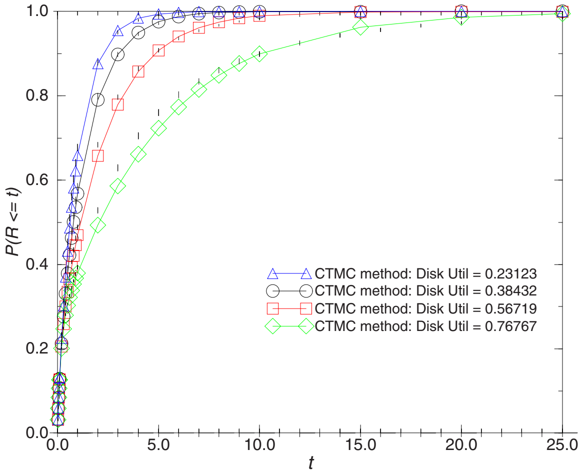


Fig. 14.22: Response time distribution in the CPU and disk queuing system

open queuing networks in which the service times and arrival processes are non-Markovian [28]. For doing so, we make use of several existing results on the response time distribution at a single queue. Using these, a queuing network is translated into a Markov or semi-Markov chain, whose absorption time distribution approximates the response time distribution of the queuing network.

In Sect. 14.5.1, we consider the approximation of the response time distribution of a network of queues containing $M/G/1$ priority queues. The approach is extended to $PH/G/1$ queues in Sect. 14.5.2.

14.5.1 Approximating non-PH distributions

In the previous sections we dealt with queues whose response time distribution could be expressed as the absorption time distribution of a CTMC; or in other words, their response time distribution was phase-type. However, this is true for only few types of queues. We shall now extend our approach to queues whose waiting (or response) time distribution is not phase-type. We consider a multiple class queuing network of $M/M/1$ queues with the *priority* service discipline. We start with first reviewing some results on the response time distribution of the $M/G/1$ priority queue.

14.5.1.1 The response time distribution at an $M/G/1$ priority queue

The Laplace-Stieltjes transform of the waiting time distribution at an $M/G/1$ priority queue has been derived by Takács [41]. It is assumed that the priorities of the arriving jobs are independent, identically distributed random variables, independent of the arrival times, and a job having a smaller priority number has preference over a job with a greater priority number. Some required notation is described in Table 14.1.

Symbol	Description	Definition
λ	Arrival rate of jobs	
H	Discrete random variable indicating the job priority	
p_h	Probability that arriving job has priority h $(h=1,2,\ldots)$	$P(H=h)$
$F_H(h)$	Probability that arriving job has priority $\leq h$	$P(H\leq h)=\sum_{i=1}^{h} p_i$
$\lambda_{[1,h]}$	Arrival rate of jobs with priority $\leq h$	$\lambda F_H(h)$
B	Random variable indicating the job service time	
$F_B(t)$	Service time distribution	$P(B\leq t)$
β	Mean service time	$E(B)=\int_0^\infty t\,dF_B(t)$
$F_{B\vert H}(t\mid H=h)$	Conditional service time distribution for jobs with priority h	
$f^*_{B\vert H}(s\mid H=h)$	LST of conditional service time distribution for jobs with priority h	$\int_0^\infty \exp(-st)\,dF_{B\vert H}(t\mid H=h)$
$F_{B\vert H}(t\mid H\leq h)$	Conditional service time distribution for jobs with priority $\leq h$	$\frac{\sum_{i=1}^{h} F_{B\vert H}(t\vert H=i)\cdot p_i}{F_H(h)}$
$f^*_{B\vert H}(s\mid H\leq h)$	LST of conditional service time distribution for jobs with priority $\leq h$	$\int_0^\infty \exp(-st)\,dF_{B\vert H}(t\mid H\leq h)$
$\beta_{[1,h]}$	Conditional mean service time for jobs with priority $\leq h$	$E(B\mid H\leq h)$ $=\int_0^\infty t\,dF_{B\vert H}(t\mid H\leq h)$
q	$q=\begin{cases}\infty & \text{if } \lambda_{[1,h]}\beta_{[1,h]}<1\ \forall h,\\ \min\{h:\lambda_{[1,h]}\beta_{[1,h]}\geq 1\} & \text{otherwise}\end{cases}$	
$\lambda_{(h,q)}$	Arrival rate for jobs with priority $\in (h,q)$	$\lambda\,(F_H(q-1)-F_H(h))$
$F_{B\vert H}(t\vert h<H<q)$	Conditional service time distribution for jobs with priority $\in (h,q)$	$\frac{\sum_{i=h+1}^{q-1} F_{B\vert H}(t\vert H=i)\cdot p_i}{F_H(q-1)-F_H(h)}$
$f^*_{B\vert H}(s\vert h<H<q)$	LST of conditional service time distribution for jobs with priority $\in (h,q)$	$\int_0^\infty \exp(-st)\,dF_{B\vert H}(t\vert h<H<q)$
W	Random variable indicating the job wait time	
$F_{W\vert H}(t\mid H=h)$	Conditional wait time distribution for jobs with priority h	
$f^*_{W\vert H}(s\mid H=h)$	LST of conditional wait time distribution for jobs with priority h	$\int_0^\infty \exp(-st)\,dF_{W\vert H}(t\mid H=h)$
$F_{W\vert H}(t\mid H\leq h)$	Conditional wait time distribution for jobs with priority $\leq h$	$\frac{\sum_{i=1}^{h} F_{W\vert H}(t\vert H=i)\cdot p_i}{F_H(h)}$
$f^*_{W\vert H}(s\mid H\leq h)$	LST of conditional wait time distribution for jobs with priority $\leq h$	$\int_0^\infty \exp(-st)\,dF_{W\vert H}(t\mid H\leq h)$

Table 14.1: Notation for LST of priority queues

If $\lambda_{[1,h]}\beta_{[1,h]} < 1$, the LST of the waiting time distribution for a priority h job [41] is given by

$$f^*_{W|H}(s \mid H = h) = f^*_{W|H}(s + \lambda_{[1,h-1]}(1 - \delta_h(s)) \mid H \leq h), \qquad (14.1)$$

where $\delta_h(s)$ is the root with the smallest absolute value in z of the equation

$$z = f^*_{B|H}(s + \lambda_{[1,h-1]}(1 - z) \mid H \leq h - 1).$$

For the case of preemptive-resume priority,

$$f^*_{W|H}(s \mid H \leq h) = \frac{1 - \lambda_{[1,h]}\beta_{[1,h]}}{1 - \lambda_{[1,h]} \cdot \frac{1 - f^*_{B|H}(s|H \leq h)}{s}}, \qquad (14.2)$$

and the LST of the conditional response time distribution is

$$f^*_{R|H}(s \mid H = h) = f^*_{W|H}(s \mid H = h) \cdot f^*_{B|H}(s + \lambda_{[1,h-1]}(1 - \delta_h(s)) \mid H = h).$$

For the non-preemptive priority case, with the condition that $\lambda_{[1,h]}\beta_{[1,h]} < 1$ for every h,

$$f^*_{W|H}(s \mid H \leq h) = \frac{1 - \lambda\beta + \lambda_{[h,q]} \cdot \frac{1 - f^*_{B|H}(s|h < H < q)}{s}}{1 - \lambda_{(1,h)} \cdot \frac{1 - f^*_{B|H}(s|H \leq h)}{s}},$$

and the LST of the conditional response time distribution is

$$f^*_{R|H}(s \mid H = h) = f^*_{W|H}(s \mid H = h) \cdot f^*_{B|H}(s \mid H = h).$$

Let us consider the simplest special case of an exponential service time distribution with rate μ_h for priority class h. Suppose there are two priority classes with preemptive-resume priority. Then for priority 1 jobs Eqs. (14.1) and (14.2) simplify to

$$f^*_{W|H}(s \mid H = 1) = \frac{1 - \lambda_{[1,1]}\beta_{[1,1]}}{1 - \lambda_{[1,1]} \cdot \frac{1 - \frac{\mu_1}{\mu_1 + s}}{s}} = \frac{1 - \frac{p_1\lambda}{\mu_1}}{1 - \frac{p_1\lambda}{\mu_1 + s}},$$

which is the LST of the $M/M/1$, FCFS waiting time distribution, as expected.

According to Eq. (14.1), for jobs with priority 2, the LST of the waiting time distribution is given by

$$\begin{aligned} f^*_{W|H}(s \mid H = 2) &= f^*_{W|H}(s + \lambda_{[1,1]}(1 - \delta_2(s)) \mid H \leq 2) \\ &= f^*_{W|H}(s + p_1\lambda(1 - \delta_2(s)) \mid H \leq 2). \end{aligned}$$

Here $\delta_2(s)$ is given by the absolute value of

$$\frac{\mu_1 + s + p_1\lambda - \sqrt{s^2 + \mu_1^2 + (p_1\lambda)^2 + 2\mu_1 s + 2sp_1\lambda - 2p_1\lambda\mu_1}}{2p_1\lambda}.$$

Clearly, $f^*_{W|H}(s \mid H = 2)$ is not rational in s and therefore does not represent a phase-type distribution. Thus it follows that even for the exponential service time distribution, the response time distribution in a multi-priority queue is not phase-type. The CTMC approach outlined earlier therefore does not directly apply. One alternative approach is to fit a phase-type distribution to the response time distribution. The advantage of this method is that the problem still reduces to computation of the transient solution of a CTMC, a problem for which many different and numerically stable solutions exist.

14.5.1.2 The CTMC approach

The main problem in the CTMC approach is to compute a good phase-type fit to the response time distribution at each queue. A lot of work has been done in the area of fitting a phase-type distribution when the first few moments of a distribution are available. The most thorough work has been presented in a series of papers by Johnson and Taaffe; see the list of references in [20]. In the context of our problem we make the following remarks.

The response time distribution of the $M/M/1$ priority queue is available to us only in the form of its LST. As one alternative, we could invert the LST numerically and fit a phase-type distribution to the distribution that results from this inversion. We do not find this alternative very prudent. This is because if we must use LST inversion, the advantage of using the CTMC approach is lost. We could model the response time problem simply as a semi-Markov chain (explained in detail later in Sect. 14.5.2) and use LST inversion for its solution, without going through the phase-type approximation. Therefore, if a matching must be done, it should be without carrying out LST inversion. We would also like to point out the key difference in other works in queuing systems based on phase-type fitting [22]: The phase-type fit is in most cases computed for the *service time* distribution. The idea in that case is to make the rest of the computations tractable. In our approach, however, we “skip” over this step, and compute a direct fit to the waiting time distribution. Such an approach is possible only when a closed form expression exists for (the LST of) the waiting time distribution at the queue under consideration. This approach keeps the state space of our resulting CTMC model from getting very large.

The availability of the LST implies that not only do we have a unique representation of the distribution in s domain but also any number of its moments are immediately available to us. We therefore have two alternatives: (1) fit the moments of the distribution; (2) fit the LST of a phase-type distribution to the given LST, by function fitting procedures. We shall discuss each of these approaches next.

14.5.1.3 Moment matching

The problem of fitting phase-type distributions to a general distribution has received a lot of attention in the past few decades. In the early 1990s, Johnson and Taaffe

(J&T) [21, 22] and Johnson [20] proposed methods by which the first three moments of a general distribution could be matched with moments of a mixture of two Erlang distributions. More recently, Osogami and Harchol-Balter [35] presented an efficient closed form solution for matching the first three moments of a subset of phase-type distributions, termed "Erlang-Coxian," which results in a nearly minimal number of phases. Horváth and Telek [16] suggested a method by which any number of moments can be matched by those of an acyclic phase-type distribution. There has been further work by Horváth et al. [17] in fitting interarrival distributions with Markovian arrival processes (MAPs), including approximation of the n-lag correlation of interarrival times.

The work presented here uses the methods of fitting phase type distributions proposed by Johnson and Taaffe which involve matching moments. Although in [20] it is mentioned that their software provides the option of fitting the LST directly, there is no discussion on this issue. We therefore explore this method in some detail in Sect. 14.5.1.4. In this section, we first develop the moment matching approach. Without going into details, we shall mention some features of J&T's moment matching technique.

In this section, we use CTMC-like graphs describing the distributions, which consist of nodes and edges. Labels on the edges indicate the probability of that edge being traversed. If there is no label, the probability is assumed to be one. The node is labeled with the rate of the exponentially distributed wait time at that node.

Johnson and Taaffe [21] have proven that a mixture of two Erlang distributions (Fig. 14.23) can match the first three moments of any distribution for which there exists some phase-type distribution that matches the first three moments. These authors have also derived conditions under which a mixed Erlang matching can be made.

Using the program MEMOM, supplied by Johnson [20], one can give the first three moments as an input, and obtain the five parameters $(p, \lambda_1, \lambda_2, n_1, n_2)$ of a mixed Erlang distribution. The first three moments of the waiting time in a priority queue have been derived in [41].

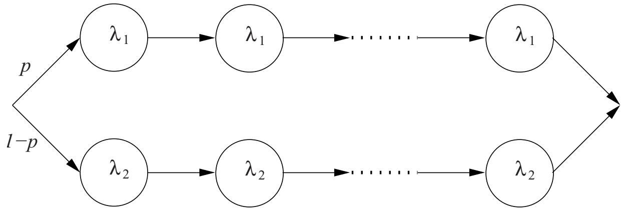


Fig. 14.23: The mixed Erlang distribution

14.5.1.4 Function fitting of the LST

The phase-type distribution that we choose for fitting is the branching Erlang distribution shown in Fig. 14.24.

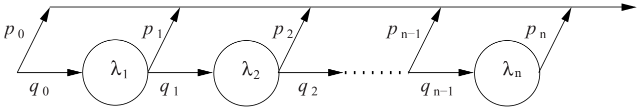


Fig. 14.24: The branching Erlang distribution

Note that the probabilities p_i and q_i add up to one for all $i = 0, 1, ...$ Since there is only one edge leaving the node with value λ_n, $p_n = 1$. Let M denote a random variable following an n-stage branching Erlang distribution. The LST of the distribution of M is

$$f_M^*(s) = p_0 + \sum_{i=1}^{n} p_i \prod_{j=1}^{i} q_{j-1} \frac{\lambda_j}{\lambda_j + s}. \tag{14.3}$$

The parameters of $f_M^*(s)$ must be chosen such that it best approximates $f_W^*(s)$. To do this, we "discretize" the problem; i.e., we appropriately choose k points $s_1, s_2, \ldots, s_k$ and minimize the function

$$\sum_{i=1}^{k} (f_M^*(s_i) - f_W^*(s_i))^2 \tag{14.4}$$

with respect to the parameters of the branching Erlang distribution. This will give the least squares fit of the function. The issues now are (1) to restrict the branching Erlang, and (2) to appropriately choose the discretization points.

The branching Erlang can be restricted to the condition $\lambda_1 = \lambda_2 = \ldots = \lambda_n =: \lambda$. We also restrict n to 3. We have found that we obtain very good fits in most cases under these restrictions. In case a good fit is not found, we increase the number of stages in the branching Erlang distribution. Furthermore, one of the parameters, p_0, is already determined to be identical to $\lim_{s\to\infty} f_W^*(s)$, which is the probability that the waiting time is zero.

We can also make this a more hybrid fitting by determining one more parameter based on matching the mean of the distribution. Suppose μ is the mean waiting time. If we choose λ by matching the means of the two distributions, it is given by

$$\lambda = \frac{q_0 p_1}{\mu} + \frac{2 q_0 q_1 p_2}{\mu} + \frac{3 q_0 q_1 q_2}{\mu}.$$

The only parameters to be chosen now are p_1 and p_2.

The choice of the points $s_1, s_2, \ldots$ significantly affects the approximation error. If we want the value of the sum in Eq. (14.4) to be $\leq k\varepsilon^2$, then we must choose at least one s_i such that $|f_M^*(s_i)| > \varepsilon/2$ or $|f_W^*(s_i)| > \varepsilon/2$. This is because if both $|f_M^*(s_i)| < \varepsilon/2$ and $|f_W^*(s_i)| < \varepsilon/2$ then $|f_M^*(s_i) - f_W^*(s_i)| < \varepsilon$, for *any* choice of parameters. The least squares fit over the discretized function will therefore not result in a good fit for the actual function. In practice, we choose points $s_1, s_2, \ldots$ such that they "contribute" largely towards the error. Thus we choose evenly spaced real values of s_i's in the interval $(0, s_k]$ such that $|f_W^*(s_k)| > \varepsilon/2$.

14.5.1.5 Example: Transaction processing system

Consider a transaction processing system maintaining information which is regularly read and updated on two databases (on DISK1 and DISK2); see Fig. 14.25(a). We would like to provide the read tasks with as up-to-date information as possible. One way to achieve this effect is to give preference to the update tasks, so that the read tasks are executed after the latest update has been performed. Thus update tasks are assigned higher priority than the read tasks. To avoid excessive scheduling overhead, the system adopts non-preemptive priority at the front-end processor.

Both read and update tasks use the processor for an amount of time that follows an exponential distribution with rate μ_{C1}. The tasks then proceed to disk D_1 with probability p_1 and to another disk D_2 with probability p_2. The service time at disk D_1 is exponentially distribution with rate μ_{D1}, and that at disk D_2 is exponentially distributed with rate μ_{D2}, for both kinds of tasks. The response time distribution for each of these tasks can be computed using the approach outlined above.

We fit the waiting time distribution at the CPU with a 2-stage branching Erlang distribution by matching the LSTs of the two distributions at a finite number of real values of s. A CTMC model, depicted in Fig. 14.25(b), is then built whose absorption time distribution approximates the response time distribution of a customer in this queuing network. Suppose the fraction of update tasks coming to the system is 0.3 and that of read tasks is 0.7. Suppose $\mu_{D1} = 1.0$, $\mu_{D2} = 0.7$, $p_1 = 0.5$, $p_2 = 0.2$ and $\mu_{C1} = 1.0$. We show the relative percentage error between the response time distribution values computed by the CTMC method and values derived by simulation. For simulation values, we use the midpoints of 95% confidence intervals. Figures 14.26(a)–(d) show the relative percentage error for arrival rates 0.04, 0.05, 0.06, and 0.07, for priority 1 customers.

Figures 14.27(a)–(d) show the relative percentage error for arrival rates 0.04, 0.05, 0.06, and 0.07, for priority 2 customers. The CPU utilization level varies from 22% to 40%.

The same example can be solved by fitting a mixed Erlang distribution to the waiting time distribution using the J&T method. That is, we fit the first three moments of a mixed Erlang distribution to the first three moments of the waiting time distribution. Figure 14.28 shows the response time distribution for a priority 2 customer, when arrival rate is 0.07. The relative percentage error is depicted in Fig. 14.29. This experiment suggests that the moment matching approach does not work as well as the hybrid approach, in which the first moment is matched and the LST is directly fitted.

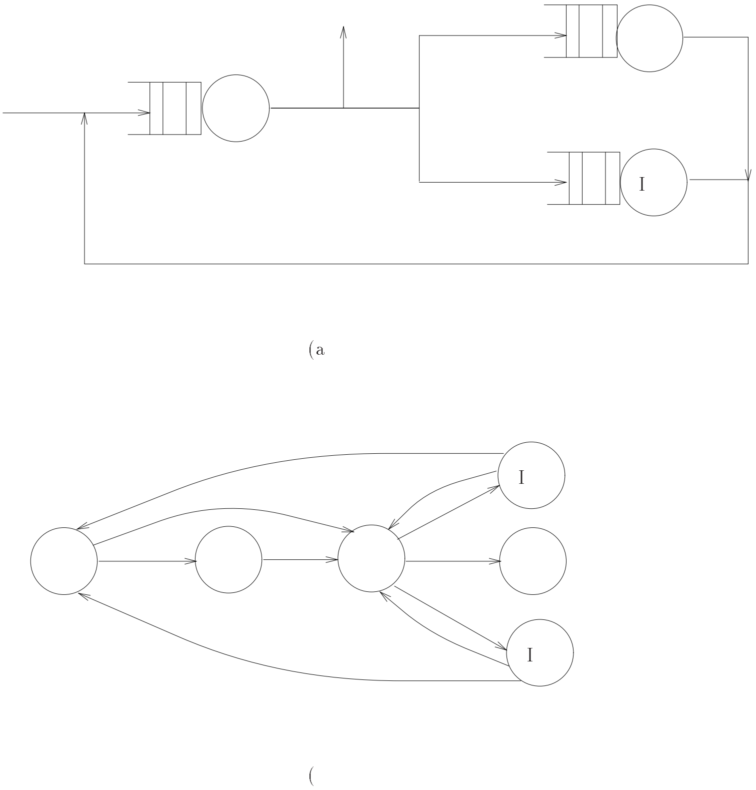


Fig. 14.25: (a) Priority queuing system (b) CTMC corresponding to the response time distribution

It is well known that the Laplace transform inversion function is unstable; i.e., small perturbations in the value of the Laplace transform $f^*(s)$ may lead to large changes in the time-domain function $f(t)$ [4]. However, it has also been noted in the literature [4] that functions that are essentially smooth are not very sensitive to perturbations in the LST.

The response time distribution functions that we use are bound to be smooth and "well-behaved," in the sense that they cannot have spikes or oscillations. Our method of approximating the LST of an unknown function with the LST of a known function thus gives good numerical results in most cases.

However, this method is disadvantageous in case a good fit for the LST cannot be found sufficiently fast. Our primary aim of a fast numerical solution is then not met. In the next section, we therefore propose a new technique, in which we do not attempt to fit a phase-type distribution. Instead, we use semi-Markov chains. This approach also extends our technique for response time computation to a network of $PH/G/1$ queues.

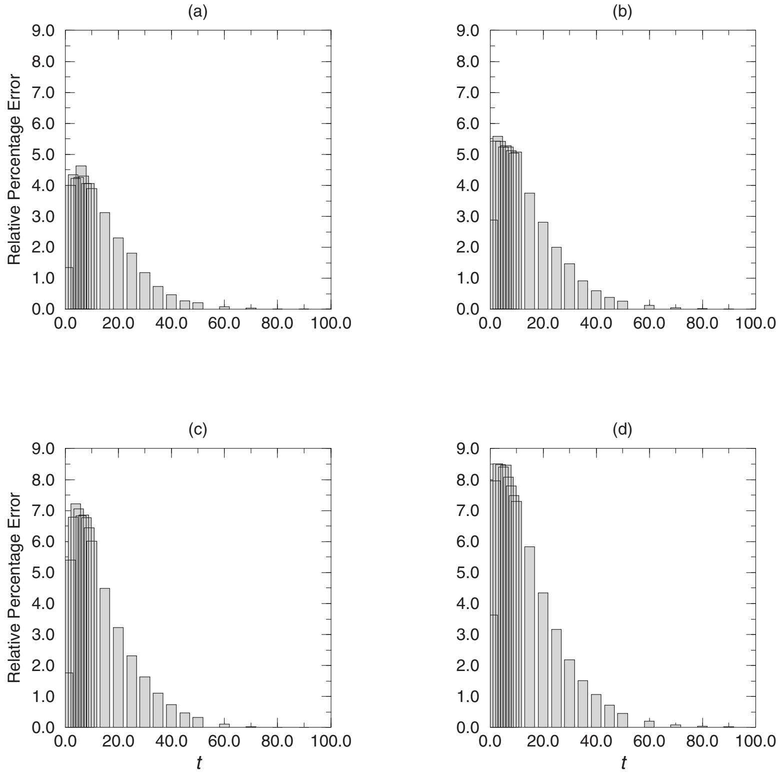


Fig. 14.26: Relative percentage error for priority 1 customers, LST fitting

14.5.2 *Modeling response time distributions using semi-Markov processes*

In the previous sections, we dealt exclusively with exponential and phase-type service time distributions and Poisson arrival processes. The simplifying assumption in our approach was that arrivals to all queues are Poisson, which is not generally the case. This is because departures from previous queues may not be Poisson (except for cases mentioned in Sect. 14.3). In this section, we address this problem and relax the Poisson arrival assumption. We allow external arrivals to the queues to be *phase-type renewal processes*, i.e., the time between arrivals is identically and independently distributed and has a PH distribution.

We extend our approach to an open network of $PH/G/1$ queues. For dealing with arrival processes that are renewal processes, we adopt techniques employed by Whitt's Queuing Network Analyzer [47], with some differences. Firstly, Whitt's approach does not require the arrival renewal processes to be phase-type. Secondly, Queuing Network Analyzer does not deal with response time distributions in detail.

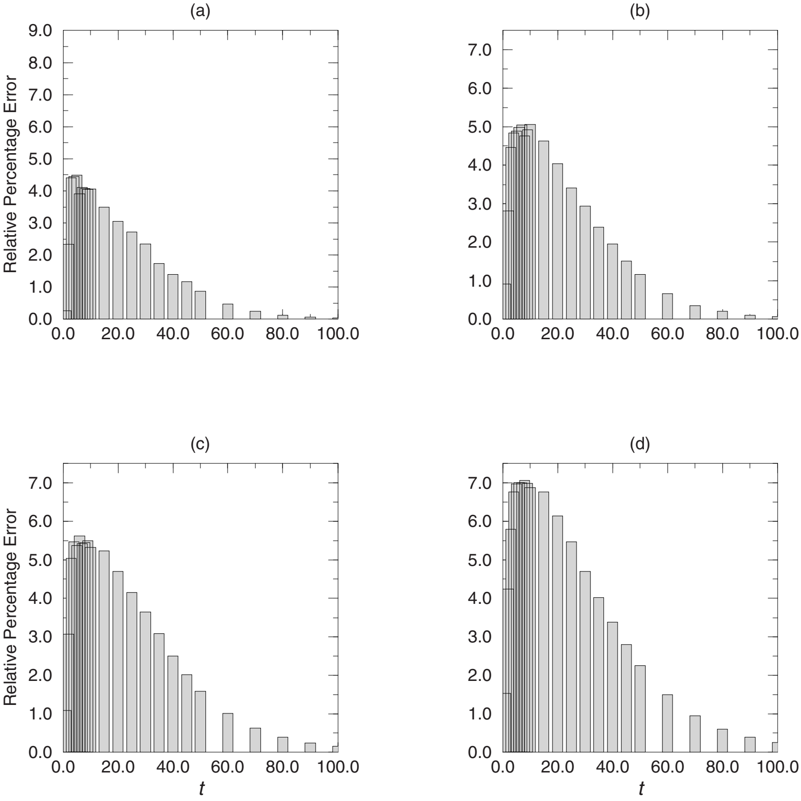


Fig. 14.27: Relative percentage error for priority 2 customers, LST fitting

By letting the arrival processes to be restricted to phase-type renewal processes, we can explore more accurate approximations to the response time distribution in the queuing network. Thirdly, unlike Whitt we do not take into account multiple server queues and the possibility of customer creation or combination.

In general, the response time distribution at a $PH/G/1$ queue is not phase-type, and hence the CTMC approach developed earlier does not directly apply. The same paradigm though can be extended to non-phase-type distributions by employing *semi-Markov processes*. For the states of such a process, the holding time distributions do not necessarily follow an exponential distribution. The future may depend on how much time has been spent in the current state. However, semi-Markov processes do maintain the "memoryless" property to the extent that all the past can be "forgotten" at a state transition epoch. This scenario lends itself very favorably to modeling response times in queuing networks. Thus for each queue in the network we create one state representing the sojourn time of a customer at that queue. Because of our independence assumptions and Markovian routing it is clear that when a customer begins sojourn at one queue, the customer history may be "forgotten."

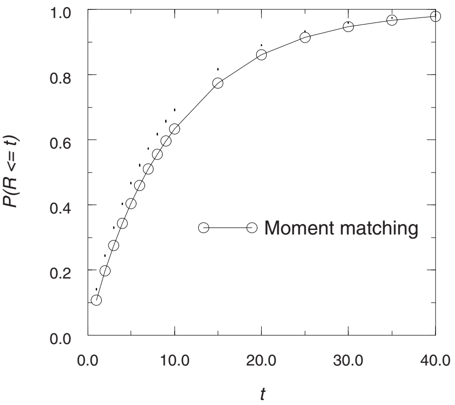


Fig. 14.28: Response time distribution for priority 2 customers, moment matching

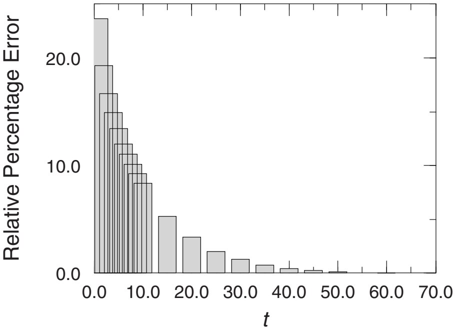


Fig. 14.29: Relative percentage error for priority 2 customers, moment matching

The sojourn time at one queue, however, is in general not exponentially distributed and hence must be "remembered." The two quantities needed for solution of a semi-Markov process are thus readily available to us: the holding times vector and the probability matrix of the embedded DTMC (which is the same as the network routing matrix).

The semi-Markov process method avoids the state space explosion that may be caused by fitting a phase-type distribution to a general distribution. Fitting also introduces an approximation error. In the semi-Markov approach we avoid these problems at the cost of a less efficient solution method (namely, numerical inversion of an LST).

There are many issues to be addressed: (1) deriving the parameters of the phase-type renewal process to each queue, (2) computing the response time distribution at each queue, (3) computing the response time distribution in the queuing network.

In the following subsections we shall explain the techniques that we developed or chose from literature to address each one of these issues.

14.5.2.1 Deriving parameters of arrival processes

For using our basic approach of decomposition of queues we need to view each queue as an isolated $PH/G/1$ queue. Therefore, we must derive the parameters of the (approximated) PH renewal arrival process to each queue. For this task, we essentially adopt Whitt's approach of characterizing the renewal processes by two parameters, namely, the mean and the squared coefficient of variation of the interarrival times. We then fit a phase-type renewal process to these two moments, and analyze each queue as a $PH/G/1$ queue. A similar approach was used by Haverkort [15] to solve a network of queues with PH service times using exact analysis. However, the problem of response time distribution was not addressed in that work.

Consider a network of m queues, $1,2,\ldots,m$. The arrival rate to each queue is computed in exactly the same way as was done under the Poisson interarrival assumption, see Eq. (14.3). In the following, the notation that was defined in Sect. 14.3.2 still holds.

The coefficients of variation are computed according to equations derived by Whitt. We reproduce them below without explanation. (For details as to how these are derived, the reader is referred to [47].)

Let the arrival rate to queue j from queue i be

$$\lambda_{ij} = \lambda_i r_{ij},$$

and the proportion of arrivals to j that came from i, $i \geq 0$,

$$f_{ij} = \lambda_{ij}/\lambda_j.$$

Let c^2_{aj} denote the squared coefficient of variation of the effective arrival process to queue j. Let c^2_{0j} denote the squared coefficient of variation for the external arrival process to queue j. Let c^2_{sj} denote the squared coefficient of variation of the service time distribution at queue j. Then the squared coefficient of the effective arrival process at queue j is obtained by solving the linear system

$$c^2_{aj} = a_j + \sum_{i=1}^{m} c^2_{ai} b_{ij}, \quad 1 \leq j \leq m, \tag{14.5}$$

where a_j and b_{ij} are constants computed as

$$a_j = 1 + w_j \left\{ (f_{0j} c^2_{0j} - 1) + \sum_{i=1}^{m} f_{ij} [(1 - r_{ij}) + r_{ij} \rho_i^2 x_i] \right\}$$

and

$$b_{ij} = w_j f_{ij} r_{ij}(1-\rho_i^2),$$

where ρ_i is the traffic intensity at queue i. Further, x_i and w_j are given by

$$x_i = \max\{c_{si}^2, 0.2\}$$

and

$$w_j = [1+4(1-\rho_j)^2(\nu_j - 1)]^{-1},$$

where

$$\nu_j = \left(\sum_{i=0}^{m} f_{ij}^2\right)^{-1}.$$

Thus, λ_i and c_{ai}^2 give us the average rate and the squared coefficient of variation of the interarrival time at queue i.

Once we have solved the two systems of linear equations, Eqs. (14.3) and (14.5), we must fit approximate PH renewal arrival processes to these "derived" parameters. To this end, we employ an adapted version of Whitt's approach [47]. (Note that this approach was used by Whitt for approximating the delay distribution at each of the queues in the network.)

Case 1: $c_{aj}^2 \geq 1.01$.
Then let the PH distribution be hyperexponential with two stages with rates γ_1 and γ_2, respectively, where the stage with rate γ_1 is chosen with probability

$$p = \frac{1 + \sqrt{(c_{aj}^2 - 1)/(c_{aj}^2 + 1)}}{2},$$

and the rates are given by

$$\gamma_1 = 2p\lambda_j \text{ and } \gamma_2 = 2(1-p)\lambda_j.$$

Case 2: $0.99 \leq c_{aj}^2 \leq 1.01$.
Let the interarrival distribution be exponential with rate λ_j.

Case 3: $0.501 \leq c_{aj}^2 \leq 0.99$.
The interarrival time distribution is assumed to be hypoexponential; i.e., it is a convolution of two exponential stages with parameters γ_1 and γ_2, respectively, where

$$\gamma_2^{-1} = \frac{\lambda_j^{-1} + \sqrt{2\lambda_j^{-2} c_{aj}^2 - \lambda_j^{-2}}}{2}$$

and

$$\gamma_1^{-1} = \lambda_j^{-1} - \gamma_2^{-1}.$$

Case 4: $c_{aj}^2 \leq 0.501$.
Let the distribution be Erlang with $k = \lceil 1/c_{aj}^2 \rceil$ stages. The rate of each stage is then $k\lambda_j$.

14.5.2.2 Response time distribution at a $PH/G/1$ queue

The expression for the LST of the waiting time distribution of a customer at a $PH/G/1$ queue may be found in Cohen's book [11]. However Cohen's method requires the computation of the $n+1$ roots of a non-linear equation (where $n+1$ is the number of states in the CTMC corresponding to the PH interarrival distribution). More recently, Lucantoni [25] developed computational algorithms for analysis of the $BMAP/G/1$ queue, which resulted in simplified algorithms for several other queues that are special cases of the $BMAP/G/1$ queue. Since the PH arrival process is a special case of the batch Markovian arrival process (BMAP), we use Lucantoni's algorithms for our computation. In the following paragraphs we shall briefly outline the computational algorithm developed by Lucantoni. Note that describing the "semantics" of the various vectors and matrices associated with Lucantoni's algorithm is beyond the scope of this chapter. We shall describe the computation in a solely mathematically complete manner. For a thorough understanding of the algorithm please refer to [25].

Suppose the PH representation of the interarrival time to a $PH/G/1$ queue is $(\alpha, \mathbf{T})$. Then the *matrix generating function* [25] associated with this interarrival time is given by $\mathbf{D}(z) = \mathbf{T} + z\tau'\alpha$. As before, $\mathbf{T}\mathbf{e}' + \tau' = \mathbf{0}'$; i.e., $\tau' = -\mathbf{T}\mathbf{e}'$. Also, ρ is again the traffic intensity, and $f_B^*(s)$ denotes the LST of the service time distribution, $F_B(t)$. Let the random variable W_v be the *virtual waiting time*, defined as the waiting time of a "virtual" customer at any arbitrary instant (or, in other words, the total "work" remaining to be done in the queue at any time), and the random variable J be the phase that the arrival process is in. The LST of the joint density function of W_v and J is denoted by $f^*_{W_v,J}(s,j)$. Then the corresponding vector $\mathbf{f}^*_{W_v}(s) = \left(f^*_{W_v,J}(s,1), f^*_{W_v,J}(s,2), \ldots, f^*_{W_v,J}(s,n+1)\right)$ is given by

$$\mathbf{f}^*_{W_v}(s) = s(1-\rho)\mathbf{g}[s\mathbf{I} + \mathbf{D}(f_B^*(s))]^{-1},$$

where $\mathbf{I}$ is an identity matrix (i.e., a matrix with a diagonal of ones and off-diagonal elements of value zero). The vector $\mathbf{g}$ is the stationary vector corresponding to $\mathbf{G}$ (a square matrix of the same size as $\mathbf{T}$), which for the PH renewal arrival process is given by

$$\mathbf{G} = \int_0^\infty \exp[(\mathbf{T} + \tau'\alpha\mathbf{G})t]\, dF_B(t).$$

Here $\mathbf{T} + \tau'\alpha\mathbf{G}$ is an infinitesimal generator matrix of a CTMC. Letting $\mathbf{u} = \alpha\mathbf{G}$, we have

$$\mathbf{u} = \int_0^\infty \alpha \exp[(\mathbf{T} + \tau'\mathbf{u})t]\, dF_B(t).$$

This matrix exponential is best computed using the uniformization [19] technique. Thus if we define $\theta = \max_i[-(\mathbf{T} + \tau'\mathbf{u})]_{ii}$ and $\gamma_n = \int_0^\infty \exp(-\theta t)\frac{(-\theta t)^n}{n!}\, dF_B(t)$, for $n \geq 0$, we can write $\mathbf{u}$ as

$$\mathbf{u} = \sum_{n=0}^{\infty} \gamma_n \alpha(\mathbf{I} + \theta^{-1}(\mathbf{T} + \tau'\mathbf{u}))^n.$$

Once **u** is computed (to a satisfactory degree of accuracy), **g** is given by

$$\mathbf{g} = \frac{\mathbf{v}}{\mathbf{v}\mathbf{e}'},$$

where **v** is the solution to the linear system $\mathbf{vT} = \mathbf{u}$.

The LST of $F_W(t)$, the actual waiting time distribution of an arbitrary customer in the queue, is given by [26]

$$f_W^*(s) = \frac{1}{\lambda(1 - f_B^*(s))} \mathbf{f}_{W_v}^*(s)[\mathbf{D}(1) - \mathbf{D}(f_B^*(s))]\mathbf{e}',$$

which in the case of a PH arrival process simplifies to

$$f_W^*(s) = \frac{1}{\lambda} \mathbf{f}_{W_v}^*(s) \tau' \alpha (1 - f_B^*(s)) \mathbf{e}'.$$

Since the response time is the sum of waiting time and service time, the LST of the response time distribution is then given by

$$f_R^*(s) = f_W^*(s) \cdot f_B^*(s). \tag{14.6}$$

14.5.2.3 Transient solution of a semi-Markov process

Once we have found the response time distribution at each queue, a semi-Markov process corresponding to the queuing network can be built. The absorption time distribution of this semi-Markov process will approximate the response time distribution in the queuing network. For this, we must carry out transient analysis of the semi-Markov process. In this section, we describe a Laplace transform method for transient analysis of a semi-Markov process [10].

We shall first introduce some notation regarding semi-Markov processes based on the paper by Ciardo et al. [10]. Suppose that $\{X(t), t \geq 0\}$ is a right-continuous semi-Markov process with state space $S \subset \mathbb{N} = \{0, 1, 2, \ldots\}$. Suppose further that the the probability that $X(t)$ will eventually reach an absorbing state is one. We denote the set of absorbing states by A, and the set of non-absorbing states by N, respectively. Let T_k be the time of the kth transition, then $T_{k+1} - T_k$ is the time spent (i.e., the holding time) in the kth visited state. Define $T_0 = 0$. If we observe this semi-Markov process at state-transition epochs, we have a discrete-time process, in fact, a DTMC. Denote this process by $Y_k = X(T_k)$, the state reached after the kth transition. Then the *kernel* of a semi-Markov process is defined as [10]

$$\mathbf{K}(t) = [K_{i,j}(t)] = [P(Y_{k+1} = j, T_{k+1} - T_k \leq t \mid Y_k = i)].$$

The transition probability matrix of the embedded DTMC is defined as [10]

$$\mathbf{E} = [E_{i,j}] = [P(Y_{k+1} = j \mid Y_k = i)] = \mathbf{K}(\infty).$$

The holding time vector is [10]

$$\mathbf{h}(t) = [h_i(t)] = [P(T_{k+1} - T_k \le t \mid Y_k = i)] = [\sum_{j \in S} K_{i,j}(t)].$$

Note that each holding time is independent of the next state.

Let $V_{i,j}(t)$ be the probability of being in state j at time t, given that the initial state was i; i.e.,

$$V_{i,j}(t) = P(X(t) = j \mid Y_0 = i).$$

Then $V_{i,j}(t)$ is given by

$$V_{i,j}(t) = I_{(i=j)} \cdot (1 - h_i(t)) + \sum_{l \in S} \int_0^t V_{l,j}(t-u)\, dK_{i,l}(u),$$

where the indicator function $I_{(i=j)}$ is equal to one if $i = j$, and zero otherwise. Now suppose that the semi-Markov process has only one absorbing state a. (If the set of absorbing states should consist of more than one state, then a can be obtained by lumping all the states in A together.) Then the conditional distribution of the time to absorption is given by $V_{i,a}(t)$. This can be found by either numerically integrating the above equation or by the LST method. Since in our case holding times are available in the LST form, we adopt the latter approach.

Let the LST of $K_{i,j}(t)$ be denoted by $\tilde{K}_{i,j}(s) = \int_0^\infty \exp(-st)\, dK_{i,j}(t)$. Partitioning $\mathbf{K}(t)$ according to the set of non-absorbing states N and the absorbing state a, we have

$$\mathbf{K}(t) = \begin{bmatrix} \mathbf{K}^{[NN]}(t) & \mathbf{K}^{[Na]}(t) \\ \mathbf{0} & K_{a,a}(t) \end{bmatrix}.$$

The LST $\tilde{\mathbf{K}}(s)$ may also be partitioned similarly:

$$\tilde{\mathbf{K}}(s) = \begin{bmatrix} \tilde{\mathbf{K}}^{[NN]}(s) & \tilde{\mathbf{K}}^{[Na]}(s) \\ \mathbf{0} & \tilde{K}_{a,a}(s) \end{bmatrix}.$$

Denote by $\tilde{\mathbf{v}}_a(s)$, the vector of LSTs of $V_{i,a}(t)$. The solution for this vector is obtained by solving the linear system [10]

$$(\mathbf{I} - \tilde{\mathbf{K}}^{[NN]}(s))\tilde{\mathbf{v}}_a(s) = \tilde{\mathbf{K}}^{[Na]}(s). \tag{14.7}$$

Thus to compute $\tilde{\mathbf{v}}_a(s)$, we must first solve the above equation, and then apply LST numerical inversion to obtain $\mathbf{v}_a(t)$, the vector of the $V_{i,a}(t)$. For details on numerical inversion of LSTs, refer to [4, 9, 18].

14.5.2.4 Building the semi-Markov chain from the queuing network

In this section we shall outline a step-by-step procedure of the semi-Markov method of computation of response time in a network of m queues without loss of customers; i.e., $p_{bi} = 0$ for all i.

Step 1: Compute the effective arrival rate λ_i at each queue i by solving Eq. (14.3), setting $p_{bi} = 0$ for all i.

Step 2: Compute the squared coefficient of variation c_{ai}^2 of the effective arrival process at queue i, from Eq. (14.5).

Step 3: Follow the procedure in Sect. 14.5.2.1 to compute the parameters of the fitted PH arrival processes.

Step 4: Given the LST of the service time distribution, $f_{B_i}^*(s)$, for each queue i of the queuing network, use Eq. (14.6) for computing the LST of the response time distribution, $f_{R_i}^*(s)$, for each queue i.

Step 5: Create a semi-Markov process with state space S which includes m states $S_1, S_2, \ldots, S_m$: one corresponding to each queue i in the network. Add an additional state S_f which denotes exit out of the queuing network. Thus $|S| = m + 1$.

Step 6: Let r_{ij} be the probability of routing from queue i to queue j in the queuing network. Then the embedded DTMC transition probability matrix $\mathbf{E}$ is given by

$$E_{S_i,S_j} = r_{ij} \quad \forall i, j \in \{1, 2, \ldots, m\},$$

$$E_{S_i,S_f} = 1 - \sum_{j=1}^{m} r_{ij} \quad \forall i \in \{1, 2, \ldots, m\}.$$

Step 7: The LST of the semi-Markov kernel is defined as

$$\tilde{K}_{S_i,S_j}(s) = f_{R_i}^*(s) \cdot E_{S_i,S_j} \quad \forall i, j \in \{1, 2, \ldots, m\},$$

$$\tilde{K}_{S_i,S_f}(s) = f_{R_i}^*(s) \cdot E_{S_i,S_f} \quad \forall i \in \{1, 2, \ldots, m\}.$$

Step 8: For the semi-Markov process related to the response time in a queuing network, the absorbing state a of Sect. 14.5.2.3 corresponds to state S_f, while the set of states N contains the states $S_1, S_2, \ldots, S_m$. Now Eq. (14.7) may be solved using a standard linear system solution method to compute $\tilde{\mathbf{v}}_{S_f}(s)$ in the s-domain, where $\tilde{\mathbf{v}}_{S_f}(s)$ denotes the vector of LSTs of $V_{S_i,S_f}(t)$. Since each $V_{S_i,S_f}(t)$ is a probability conditioned on the initial state S_i, we must compute the LST of the total unconditional probability. An incoming job joins queue i first with probability $\lambda_{0i} / \sum_{j=1}^{m} \lambda_{0j}$. Then the LST of the (approximate) response time distribution is given by $\sum_{i=1}^{m} \lambda_{0i} \tilde{V}_{S_i,S_f}(s) / \sum_{j=1}^{m} \lambda_{0j}$. This can be numerically inverted to yield our approximation to the response time distribution in the queuing network. We implemented the above algorithm and used an existing Laplace transform inversion routine [8] for our final step.

14.5.2.5 Example: Distributed system

Consider again the distributed system of Sect. 14.3.3.2. Assume that the arrival process to the terminals is a four-stage Erlang renewal process. Let the delay at the terminals be constant, and the processing time at the rest of the queues be uniformly distributed.

We can map this queuing network to the semi-Markov process depicted in Fig. 14.30, with starting state T. The labels on the arcs denote the entries of the kernel matrix (not transition rates). Let the delay at the terminals (T) be 1. Moreover, we assume that the service time is UNIFORM(1, 2) at the front-end server (F), UNIFORM(0.5, 0.7) at the communications server (C), UNIFORM(1, 3) at the database server (D), and UNIFORM(1, 2) at the general-purpose server (P).

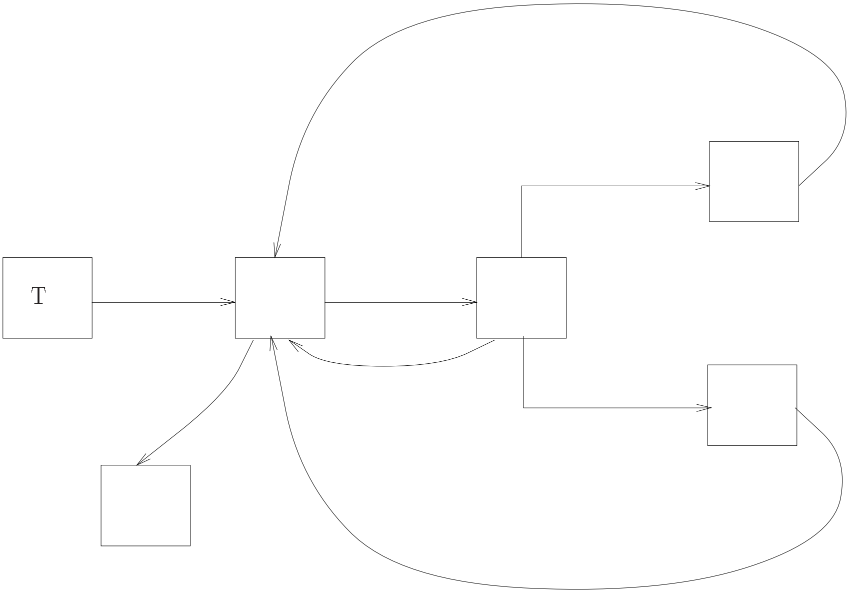


Fig. 14.30: Semi-Markov process corresponding to the response time distribution of the distributed system

Figure 14.31 shows the plot for the response time distribution for different arrival rates. The utilization at the front-end processor varied from 50% to 90%. The points used for the simulation plot for the first three values of arrival rates are the midpoints of the 99% confidence intervals, which were too narrow to be represented by vertical bars. For this example, our method required 15.6 seconds, while simulation required 2 minutes on the same machine. For an arrival rate of 0.18, the same execution time gave 95% confidence intervals of significant widths; hence the confidence intervals are shown by vertical bars.

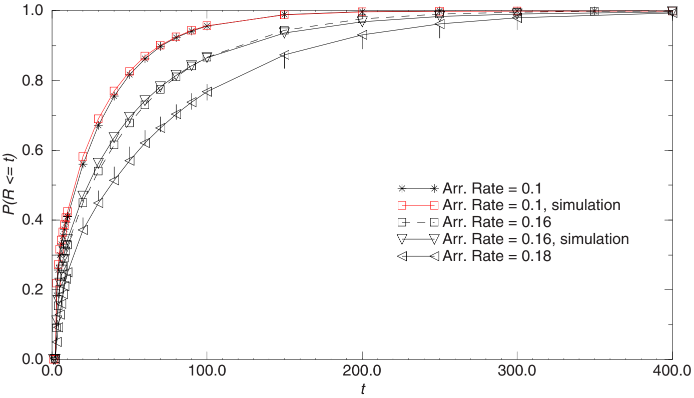


Fig. 14.31: Response time distribution in the distributed system

14.5.2.6 End-to-end delay in a virtual circuit

The building block method described in the previous sections is not the only approach in which semi-Markov processes can be employed for deriving response time distributions. In this section, we show a different way of applying semi-Markov processes, in the context of studying the end-to-end delay in a virtual circuit.

With the emergence of high-speed networks, many of the principles that governed traditional networks have undergone reevaluation. One major difference is that the propagation delay of the link is now the major contributor to the end-to-end delay of a message and not the transmission time. This affects various design choices, one instance of which is whether the error control should be end-to-end or link-by-link. It has been shown in the communications literature [5] that in the domain of high-speed networks, end-to-end error control is far superior to the link-by-link error control, when end-to-end delays are considered. Bhargava et al. [5] developed an analytical model to compute mean end-to-end delays in a virtual circuit. However, just the mean does not provide enough information about the message delay. In this section, we compute the delay distribution of a message through a virtual circuit of a high-speed network.

In the end-to-end error control scheme, the first node of a virtual circuit (VC) buffers a message until it has received an ACK from the final destination node. If an ACK is not received within a timeout period the *first* node retransmits the message. The intermediate nodes only perform error *detection*; i.e., in case the arriving message is erroneous the intermediate node simply discards the message. A message arrival to an intermediate node whose buffers are full is also lost. In both cases, a retransmission will be initiated from the first node.

The model for a four-hop virtual circuit is depicted in Fig. 14.32. Traffic from a Poisson source of rate λ_{VC} enters a virtual circuit of $m = 4$ nodes. It is assumed that all nodes except the first one are allocated b buffers. The message transmission

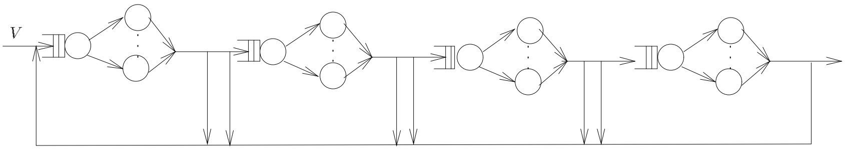


Fig. 14.32: Queuing network representing virtual circuit

rate at each node is denoted by μ and represents the effective capacity as seen by the traffic belonging to the VC under consideration. Suppose the probability of a message getting corrupted between two nodes is p. Also let q_i denote the buffer full probability at node i. Let $1/\mu_p$ denote the *fixed* propagation delay along link i. The branching after link i in Fig. 14.32 represents the retransmission of the message. Thus the message is retransmitted either if it is corrupted or if it is lost because of full buffers. The first node starts a timeout period as soon as it finishes transmission; the message is retransmitted if no ACK is received within the timeout period. The first node in this queuing network is modeled as an $M/M/1$ queue, while the others are modeled as $M/M/1/b$ queues. Let λ_i be the message arrival rate (including erroneous messages which will be discarded) to the node i. We shall compute the arrival rates using the method described in [5].

The probability q_i of having b messages at node i (≥ 2) is given by Gross et al. [13] as

$$q_i = \frac{(1-\rho_i)\rho_i^b}{1-\rho_i^{b+1}}, \quad i = 2,3,\ldots,$$

where $\rho_i = \lambda_i(1-p)/\mu$. The λ_i are computed by noting that the message throughput rate out of the network must also be λ_{VC}, and hence the message throughput at point O_m in the figure must be $\lambda_{VC}/(1-p)$ [5]. However, this should equal the message arrival rate at point I_m in the figure. Then, the following must be true [5]:

$$\lambda_m(1-p)(1-q_m) = \frac{\lambda_{VC}}{1-p}.$$

If we substitute the value of q_m in this equation, we obtain an equation in only one unknown, λ_m, which can be solved numerically. Once λ_m is computed, λ_{m-1} may be computed in a similar way, and working backwards, λ_i and q_i may be computed for all $i = 2,3,\ldots,m$.

Now, suppose N_t denotes the number of times a message must be retransmitted, before it reaches correctly to the final destination. Also let p_{fail} denote the probability that one transmission of the message fails. Let d_i denote the probability that a message is discarded by node i. Then $d_i = p+(1-p)q_i$, $i = 2,3,\ldots,m$. Further, p_{fail} is given by [5]

$$p_{fail} = \sum_{i=2}^{m} d_i \prod_{j=2}^{i-1}(1-d_j).$$

Given p_{fail}, the probability mass function of N_t is given by [5]

$$P(N_t = k) = p_{fail}^k (1 - p_{fail}), \quad k = 0, 1, \ldots$$

Let T_{ee} denote the timeout period and R_i denote the delay of the message at node *i*. Now the total delay time R required to deliver the message from the source to the final destination correctly is given by

$$R = N_t(R_1 + T_{ee}) + \sum_{i=1}^{m} (R_i + 1/\mu_p) = R_w + R_c,$$

where $R_w := N_t(R_1 + T_{ee})$ represents the time taken for all the transmissions that go wrong and $R_c := \sum_{i=1}^{m}(R_i + 1/\mu_p)$ is the time taken by the message during its final correct traversal through the virtual circuit. Then the LST of the distribution of R is given by

$$f_R^*(s) = f_{R_w}^*(s) \cdot f_{R_c}^*(s). \tag{14.8}$$

Let $R_{w1} = R_1 + T_{ee}$. Since we assume that node 1 is an $M/M/1$ queue with arrival rate λ_1, the LST of the delay distribution at this node is given by

$$f_{R_1}^*(s) = \frac{\mu - \lambda_1}{\mu - \lambda_1 + s},$$

and $f_{T_{ee}}^*(s) = \exp(-T_{ee}s)$. Then, $f_{R_{w1}}^*(s) = f_{R_1}^*(s) \cdot f_{T_{ee}}^*(s)$. Now, $f_{R_w}^*(s)$ is found by conditioning on N_t and then unconditioning:

$$\begin{aligned} f_{R_w}^*(s) &= \sum_{k=0}^{\infty} (f_{R_{w1}}^*(s))^k P(N_t = k) \\ &= \sum_{k=0}^{\infty} (f_{R_{w1}}^*(s))^k p_{fail}^k (1 - p_{fail}) \\ &= (1 - p_{fail}) \sum_{k=0}^{\infty} (f_{R_{w1}}^*(s) p_{fail})^k \\ &= \frac{1 - p_{fail}}{1 - f_{R_{w1}}^*(s) p_{fail}}. \end{aligned}$$

R_c may be represented by the absorption time distribution of a semi-Markov chain which essentially represents a tandem network. In this example, however, the LST of R_c, which we denote here by $f_{R_c}^*(s)$, may be derived in closed form simply as

$$f_{R_c}^*(s) = \prod_{i=1}^{m} \left(f_{R_i}^*(s) \exp\left(-\frac{s}{\mu_p} \right) \right). \tag{14.9}$$

The distribution of R_i is the conditional distribution of the delay at an $M/M/1/b$ queue, given that the arriving job is not lost. Its LST is given by

$$f^*_{R_i}(s) = \sum_{j=0}^{b-1} \frac{1-\rho_i}{1-\rho_i^b}\rho_i^j \left(\frac{\mu}{\mu+s}\right)^{j+1} .$$

Substituting this expression into Eq. (14.9), we are able to compute $f^*_{R_c}(s)$. The LST of the delay distribution can then be computed from Eq. (14.8). Numerical inversion of this LST gives us an approximation to the end-to-end delay distribution of a message in the virtual circuit.

Figures 14.33–14.35 show the delay distribution for various buffer sizes and arrival rates.

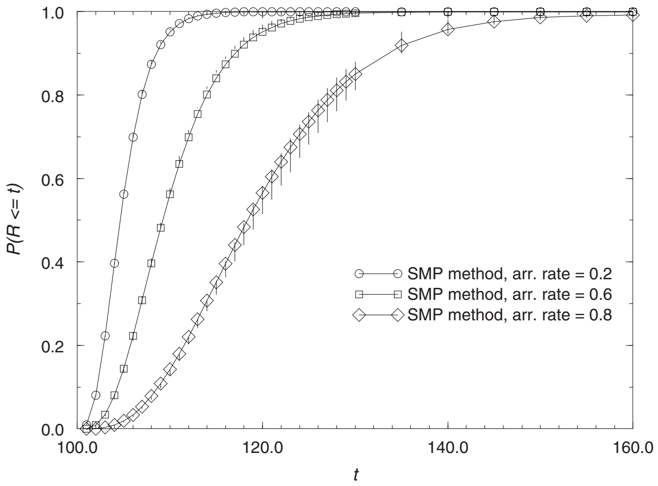


Fig. 14.33: Delay distribution in the VC, $b = 20$ buffers

In this example, we fixed the simulation time to be a maximum of 5 minutes, to study the effect on the confidence intervals. For each of the plots shown the simulation ran up to its limit of 5 minutes. The vertical bars next to the curves represent 95% confidence intervals obtained by simulation. As can be seen, for an arrival rate of 0.8, the confidence intervals become very wide. The numerical method took 42.6 seconds on the same machine.

14.6 Conclusions

In this chapter, we described three methods for computation of the response time distribution in open queuing networks. First, when the network is of queues whose response time distributions are phase-type, we presented a method to directly map the "response time building blocks" to a CTMC with absorbing states. Second, when

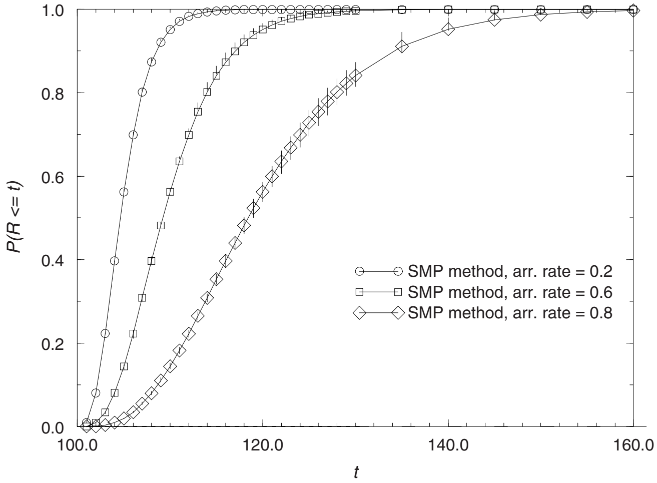


Fig. 14.34: Delay distribution in the VC, $b = 40$ buffers

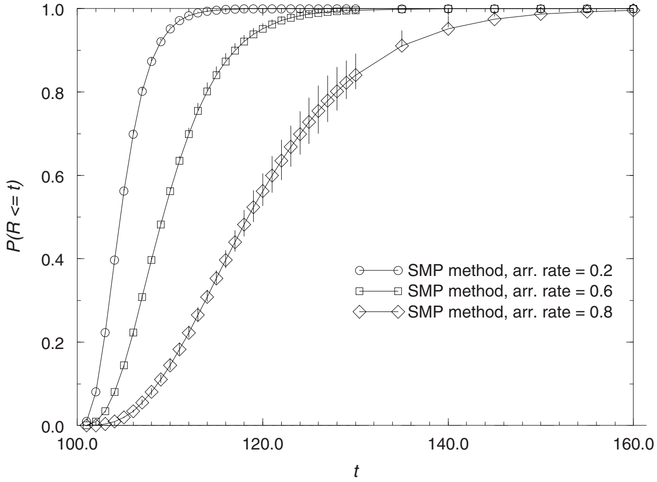


Fig. 14.35: Delay distribution in the VC, $b = 60$ buffers

the network is of queues whose response times are not phase-type, we discussed two methods of approximating the response time distributions with phase-type distributions, and then mapping the response time to the absorption time in a CTMC. Third, again for networks with queues whose response times are not phase-type, we presented an approach where the response time was computed as the time to absorption in a semi-Markov chain.

After extensive experimentation and comparison with discrete-event simulation the following general remarks can be made about the three proposed methods:

- In all cases the methods work very well for low utilization, and start degrading in accuracy at higher levels of utilization.
- The LST fitting method holds least promise because the fitting process can take a prohibitively large amount of time. This time is not justified for an approximate method.
- The semi-Markov process method works very well when the service time distribution does not have a very low coefficient of variation. Thus it does not work very well for deterministic service times.

The problem of deriving the response time distribution in queuing networks has been addressed in many different ways in the queuing network literature. For the most part, research has focused on specific cases or on solving a simpler problem such as response time through a single path in a queuing network. In this chapter, we adopted a different approach; we have attempted to solve a more general problem, at the cost of making some (judicious) approximations. The motivation was to provide a fast (but approximate) solution for a problem that otherwise can be solved only by simulation, which at times can be tiresome. The approach taken for the solution of the problem was to make use of "building blocks" that have been developed and putting them together with various "tools" to form one whole approximation method. The building blocks that our method relies on are the results on the waiting time distribution at various different kinds of queues. The tools that we used were linear system solution, fixed point iteration, phase-type fitting, transient CTMC solution and transient semi-Markov solution. Empirical studies of our method and comparison with discrete event simulation demonstrate that our method provides fast and fairly accurate predictions of response time distribution. The method is thus an alternative to simulation in providing fast answers to what-if questions regarding design issues that may affect sojourn time.

Immediate improvements possible to this method are in fitting arrival processes which also incorporate correlation between arrivals, such as the batch Markovian arrival process. As mentioned, the method also does not work well when service times are deterministic; this problem needs to be addressed. The solution method should also be extended to incorporate finite capacity queues with general arrivals and service times.

References

1. Abate, J., Choudhury, G.L., Whitt, W.: Exponential approximations for tail probabilities in queues II: Sojourn time and workload. Operations Research **44**(5), 758–763 (1996)
2. Au-Yeung, S.W., Dingle, N.J., Knottenbelt, W.J.: Efficient approximation of response time densities and quantiles in stochastic models. In: Proc. 4th International Workshop on Software and Performance, pp. 151–155 (2004)

3. Avritzer, A., Bondi, A., Grottke, M., Trivedi, K.S., Weyuker, E.J.: Performance assurance via software rejuvenation: Monitoring, statistics and algorithms. In: Proc. International Conference on Dependable Systems and Networks 2006, pp. 435–444 (2006)
4. Bellman, R.E., Kalaba, R.E., Lockett, J.A.: Numerical Inversion of the Laplace Transform. Elsevier, New York (1966)
5. Bhargava, A., Kurose, J.F., Towsley, D., Vanleemput, G.: Performance comparison of error control schemes in high-speed computer communication networks. IEEE Journal on Selected Areas in Communications **6**(9), 1565–1575 (1988)
6. Boxma, O.J., Daduna, H.: Sojourn times in queueing networks. In: H. Takagi (ed.) Stochastic Analysis of Computer and Communication Systems, pp. 401–450. Elsevier, Amsterdam (1990)
7. Bradley, J.T., Dingle, N.J., Knottenbelt, W.J., Wilson, H.J.: Hypergraph-based parallel computation of passage time densities in large semi-Markov models. In: Proc. 4th International Workshop on Numerical Solution of Markov Chains, pp. 99–120 (2003)
8. Chimento, P.F.: System performance in a failure prone environment. Ph.D. thesis, Department of Computer Science, Duke University, Durham (1988)
9. Chimento, P.F., Trivedi, K.S.: The completion time of programs on processors subject to failure and repair. IEEE Transactions on Computers **42**(10), 1184–1194 (1993)
10. Ciardo, G., Marie, R.A., Sericola, B., Trivedi, K.S.: Performability analysis using semi-Markov reward processes. IEEE Transactions on Computers **39**(10), 1251–1264 (1990)
11. Cohen, J.W.: The Single Server Queue, 2nd edn. North-Holland, New York (1982)
12. Dingle, N.J., Harrison, P.G., Knottenbelt, W.J.: Uniformization and hypergraph partitioning for the distributed computation of response time densities in very large Markov models. Journal of Parallel and Distributed Computing **64**(8), 908–920 (2004)
13. Gross, D., Shortle, J.F., Thompson, J.M., Harris, C.M.: Fundamentals of Queueing Theory, 4th edn. John Wiley & Sons, Hoboken (2008)
14. Harrison, P.G.: Approximate analysis and prediction of time delay distributions in networks of queues. Computer Performance **2**, 124–135 (1981)
15. Haverkort, B.R.: Approximate analysis of networks of $PH/PH/1/K$ queues: Theory & tool support. In: H. Beilner, F. Bause (eds.) MMB '95: Proceedings of the 8th International Conference on Modelling Techniques and Tools for Computer Performance Evaluation: Quantitative Evaluation of Computing and Communication Systems, *Lecture Notes in Computer Science*, vol. 977, pp. 239–253. Springer, London (1995)
16. Horváth, A., Telek, M.: Matching more than three moments with acyclic phase type distributions. Stochastic Models **23**(2), 167–194 (2007)
17. Horváth, G., Buchholz, P., Telek, M.: A MAP fitting approach with independent approximation of the inter-arrival time distribution and the lag-correlation. In: Proc. 2nd International Conference on Quantitative Evaluation of Systems, pp. 124–133 (2005)
18. Jagerman, J.R.: An inversion technique for the Laplace transform. The Bell System Technical Journal **61**(8), 1995–2002 (1982)
19. Jensen, A.: Markoff chains as an aid in the study of Markoff processes. Skandinavisk Aktuarietidsskrift **36**, 87–91 (1953)
20. Johnson, M.A.: Selecting parameters of phase distributions: Combining nonlinear programming, heuristics, and Erlang distributions. ORSA Journal on Computing **5**(1), 69–83 (1993)
21. Johnson, M.A., Taaffe, M.R.: Matching moments to phase distributions: mixtures of Erlang distributions of common order. Stochastic Models **5**(4), 711–743 (1989)
22. Johnson, M.A., Taaffe, M.R.: An investigation of phase-distribution moment-matching algorithms for use in queueing models. Queueing Systems **8**(2), 129–147 (1991)
23. Lavenberg, S.S., Reiser, M.: Stationary state probabilities of arrival instants for closed queueing networks with multiple types of customers. Journal of Applied Probability **17**(4), 1048–1061 (1980)
24. Lemoine, A.J.: Networks of queues - a survey of equilibruim analysis. Management Science **24**(4), 464–481 (1977)
25. Lucantoni, D.M.: New results on the single server queue with a batch Markovian arrival process. Stochastic Models **7**(1), 1–46 (1991)

26. Lucantoni, D.M.: The $BMAP/G/1$ queue: A tutorial. In: L. Donatiello, R. Nelson (eds.) Performance Evaluation of Computer and Communication Systems: Joint Tutorial Papers of Performance '93 and Sigmetrics '93, *Lecture Notes in Computer Science*, vol. 729, pp. 330–358. Springer, Berlin (1993)
27. MacNair, E.A., Sauer, C.H.: Elements of Practical Performance Modeling. Prentice Hall, Englewood Cliffs (1985)
28. Mainkar, V.: Solutions of large and non-markovian performance models. Ph.D. thesis, Department of Computer Science, Duke University, Durham (1994)
29. Mainkar, V., Woolet, S., Trivedi, K.S.: Fast approximate computation of response time distribution in open Markovian network of queues. In: Proc. 17th International Conference on Modelling Techniques and Tools for Computer Performance Evaluation, pp. 67–70 (1994)
30. Melamed, B., Yadin, M.: Numerical computation of sojourn-time distributions in queueing networks. Journal of the ACM **31**(4), 839–854 (1984)
31. Melamed, B., Yadin, M.: Randomization procedures in the computation of cumulative-time distributions over discrete state Markov processes. Operations Research **32**(4), 926–944 (1984)
32. Muppala, J.K.: Performance and dependability modeling using stochastic reward nets. Ph.D. thesis, Department of Electrical Engineering, Duke University, Durham (1991)
33. Muppala, J.K., Trivedi, K.S., Mainkar, V., Kulkarni, V.G.: Numerical computation of response time distributions using stochastic reward nets. Annals of Operations Research **48**(2), 155–184 (1994)
34. Neuts, M.F.: Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach, corrected edn. Dover Publications, Mineola (1994)
35. Osogami, T., Harchol-Balter, M.: Closed form solutions for mapping general distributions to quasi-minimal PH distributions. Performance Evaluation **63**, 524–552 (2006)
36. Perros, H.G.: Approximation algorithms for open queueing networks with blocking. In: H. Takagi (ed.) Stochastic Analysis of Computer and Communication Systems, pp. 451–498. Elsevier, Amsterdam (1990)
37. Sahner, R.A., Trivedi, K.S., Puliafito, A.: Performance and Reliability Analysis of Computer Systems: An Example-based Approach Using SHARPE. Kluwer Academic Publishers, Boston (1996)
38. Schassberger, R., Daduna, H.: Sojourn times in queuing networks with multiserver modes. Journal of Applied Probability **24**(2), 511–521 (1987)
39. Sevcik, K.C., Mitrani, I.: The distribution of queueing network states at input and output instants. Journal of the ACM **28**(2), 358–371 (1981)
40. Shanthikumar, J.G., Buzacott, J.A.: The time spent in a dynamic job shop. European Journal of Operational Research **17**(2), 215–226 (1984)
41. Takács, L.: Priority queues. Operations Research **12**(1), 63–74 (1964)
42. Trivedi, K.S.: Probability and Statistics with Reliability, Queuing and Computer Science Applications, 2nd edn. John Wiley & Sons, New York (2002)
43. Trivedi, K.S., Sathaye, A.S., Ibe, O.C., Howe, R.C.: Should I add a processor? In: Proc. 23rd Annual Hawaii International Conference on System Sciences, pp. 214–221 (1990)
44. Van Houdt, B., Blondia, C.: Approximated transient queue length and waiting time distributions via steady state analysis. Stochastic Models **21**(2-3), 725–744 (2005)
45. Van Velthoven, J., Van Houdt, B., Blondia, C.: Response time distribution in a $D\text{-}MAP/PH/1$ queue with general customer impatience. Stochastic Models **21**(2-3), 745–765 (2005)
46. Walrand, J., Varaiya, P.: Sojourn times and the overtaking condition in Jacksonian networks. Advances in Applied Probability **12**(4), 1000–1018 (1980)
47. Whitt, W.: The queueing network analyzer. The Bell System Technical Journal **62**(9), 2779–2815 (1983)
48. Woolet, S.P.: Performance analysis of computer networks. Ph.D. thesis, Department of Electrical Engineering, Duke University, Durham (1993)

Chapter 15
Decomposition-Based Queueing Network Analysis with FiFiQueues

Ramin Sadre, Boudewijn R. Haverkort

Abstract In this chapter we present an overview of decomposition-based analysis techniques for large open queueing networks. We present a general decomposition-based solution framework, without referring to any particular model class, and propose a general fixed-point iterative solution method for it. We concretize this framework by describing the well-known QNA method, as proposed by Whitt in the early 1980s, in that context, before describing our FiFiQueues approach. FiFiQueues allows for the efficient analysis of large open queueing networks of which the interarrival and service time distributions are of phase-type; individual queues, all with single servers, can have bounded or unbounded buffers. Next to an extensive evaluation with generally very favorable results for FiFiQueues, we also present a theorem on the existence of a fixed-point solution for FiFiQueues.

15.1 Introduction

In this chapter we present an overview of the FiFiQueues method (and supporting tool) to evaluate large open queueing networks with non-Poissonian traffic streams

Ramin Sadre
Centre for Telematics and Information Technology
Faculty of Electrical Engineering, Mathematics and Computer Science
P.O. Box 217, 7500 AE Enschede, The Netherlands,
e-mail: r.sadre@utwente.nl

Boudewijn R. Haverkort
Centre for Telematics and Information Technology
Faculty of Electrical Engineering, Mathematics and Computer Science
P.O. Box 217, 7500 AE Enschede, The Netherlands,
e-mail: b.r.h.m.haverkort@utwente.nl;
Embedded Systems Institute
P.O. Box 513, 5600 MB Eindhoven, Netherlands,
e-mail: boudewijn.haverkort@esi.nl

R.J. Boucherie and N.M. van Dijk (eds.), *Queueing Networks: A Fundamental Approach*,
International Series in Operations Research & Management Science 154,
DOI 10.1007/978-1-4419-6472-4_15, © Springer Science+Business Media, LLC 2011

and non-exponential services. FiFiQueues is an example of a decomposition-based queueing network evaluation approach, in which the overall evaluation is broken into per-queue evaluations, thus making the method highly scalable.

The earliest work on open networks of queues has probably been reported by Jackson [20]. The so-called Jackson queueing networks (JQNs) allow for the analysis of open networks of M|M|1 queues, in which jobs are routed according to fixed probabilities. The external arrival process forms a Poisson process; arrivals may be spread over more than one queue. Departures from the queueing network are also possible.

In the mid 1970s, Kühn developed an approximate evaluation approach for an extended class of models [27], including non-Poissonian arrivals, as well as service times that followed other than exponential distributions. As an extension of this approach, Whitt proposed the QNA method in the early 1980s [49, 50]; QNA can be seen as a full-fledged approach to evaluate networks of G|G|1 queues approximately. Since our FiFiQueues approach can be regarded as an extension of QNA, and still relies on some of the assumptions made in QNA, we concisely present QNA in Section 15.3.

We are not the only researchers who have worked on extensions of QNA, nor are the extensions of QNA that we describe here the only possible extensions. Schuba et al. [46] reported on work involving the inclusion of multicast communication using routing trees (instead of the usual routing chains). Heindl et al. proposed decomposition-based analysis techniques taking into account correlations in the traffic streams between the queueing nodes, e.g., by using MAPs and MMPPs as traffic descriptors, cf. [17, 19, 20, 21]. Kim et al. [24] proposed an extension of QNA to include correlations in the traffic streams. For two small networks that are studied in detail (with 2 and 3 nodes resp.) better results than with standard QNA are obtained. The question how well the method scales to larger and more complex queueing networks remains open. Finally, in 1990 Harrison and Nguyen proposed the QNET approach [12] which, however, appears impractical for large queueing networks. A simplification of QNET, called ΠNET (described in the same paper) appears more practical; however, its approach is very similar to that of (standard) QNA.

The aim of this chapter is to present in detail the complete set of extensions we have proposed, and that led to the approach now known as FiFiQueues. FiFiQueues extends QNA in two ways: first, it extends the model class, and secondly, it removes a number of approximation steps from it. In particular, we do not address general G|G|1 queues, but allow instead for both PH|PH|1 as well as PH|PH|1|K queues. That is, we allow for phase-type distributions as inter-arrival and service-times, but at the same time also allow for finite- and infinite-buffer queues. This choice has two implications. The restriction to phase-type distributions allows us to use exact analysis algorithms for the per-queue evaluations, e.g., based on matrix-geometric methods. Secondly, the introduction of finite queues allows us to model queueing networks with losses, which has a severe impact on the solution of the traffic equations and forces us to follow a fixed-point iterative algorithm to solve them.

The chapter is further organized as follows. In Section 15.3 we summarize the QNA method. In Section 15.4 we present the FiFiQueues algorithm and its implementation in an integrated tool. We then present a large number of cases to validate both basic FiFiQueues and its extensions (against simulation results) in Section 15.5. Finally, Section 15.6 presents some conclusions. To keep the chapter self-contained, we added appendices on Jackson queueing networks (Section 15.7), on Markovian arrival processes, phase-type distributions and quasi-birth-death processes (Section 15.8), as well as a proof of the existence of a fixed point for the models we study (Section 15.9).

We finally remark that we already worked on some further extensions on FiFiQueues. We extended our approach to also deal with closed queueing networks, as published in [44]. Furthermore, we developed extensions that deal with correlations in traffic streams, as well as with higher moments [40].

15.2 The decomposition approach

Sketch of the idea

A common approach to evaluate the performance of communication systems is to construct and analyze a large monolithic model, often via an underlying state-space-based representation (typically a Markov chain). However, analysis methods relying on an analysis of such a large state space usually suffer from the state space explosion phenomenon: If two models A and B with a resp. b states are composed to a new "product model" $A \times B$, this model has potentially $a \cdot b$ states (this assumes that there are no mutually exclusive states). For large systems models, the number of states quickly grows beyond what can be practically handled.

The decomposition approach aims to reduce the complexity of the analysis by decomposing the system into smaller components that are analyzed more or less independently, thus avoiding the analysis of the overall full state space. The basic idea is the following: If we have two submodels A and B, with a resp. b possible states, we avoid to construct and analyze the full "product model" $A \times B$. Instead we do the following:

1. We assume that the system has the structure $B(A)$ instead of $A \times B$, i.e., that in the resulting composition the submodel B depends on A but not vice versa.
2. Based on that assumption, we analyze model A independently of B and summarize its behavior in some so-called *descriptor* d_A.
3. The descriptor d_A is used to parameterize model B and we analyze the new model $B(d_A)$ with $b(d_A)$ states instead of $B(A)$ with $a \cdot b$ states. Hence, the decomposition approach reduces the number of states to analyze from $a \cdot b$ to $a + b(d_A)$.
4. Now, we know the behavior of A combined with B. The global behavior of the system can then be derived.

Of course, this approach only makes sense if $B(d_A)$ has fewer states than $B(A)$. In general, this requires that the descriptor d_A is only an approximation to A and, hence, the decomposition approach only provides an approximation to the original model.

In the context of queueing networks, the submodels are naturally equivalent to the individual queueing stations and the descriptor represents the inter-station traffic. For example, in a tandem queueing network with two stations A and B, the descriptor d_A is obtained by analyzing station A and actually is a description of the traffic stream that departs from station A and arrives at station B. Hence, we will often call d_A a *traffic descriptor* in the following.

Open questions

The approach described above leaves several open questions:

1. What do the traffic descriptors look like?
2. How are more complex systems analyzed? Note that in the example above, the assumed structure $B(A)$ would basically restrict the analysis to tandem queueing networks.
3. How are the individual stations analyzed?

These questions are addressed by various decomposition-based analysis methods in different ways, thus leading to different model classes. When we describe the Queueing Network Analyzer (Section 15.3), FiFiQueues (Section 15.4), and the analysis of Jackson queueing networks (Section 15.7), we have to address these three questions for each method separately. However, since we focus on the analysis of open queueing networks with feedback, we can already give some general answers to question 2 and 3 which are true for all methods.

The analysis of complex networks

In an open queueing network with N queueing stations, the traffic descriptor $desc_{i,j}$ describes the traffic stream from queueing station i to station j, with $1 \leq i, j \leq N$. The outside world is represented by a "virtual" station *ext*, hence, we denote the traffic arriving from outside to station i as $desc_{ext,i}$, and the traffic leaving the network from i as $desc_{i,ext}$. We rely on the fixed-point iteration algorithm presented in [14, 48] to analyze such networks:

```
 1 initialize all traffic descriptors $desc_{i,j}^{(0)}$:
 2       set $desc_{ij}^{(0)}$ to the *null* value if $i \neq ext$
 3       set $desc_{ij}^{(0)}$ to the specified value if $i = ext$
 4 $n := 0$
 5 **do**
 6       $n := n+1$
 7       analyze each queueing station $i$ with arrival traffic $desc_{k,i}^{(n)}$, $1 \leq k \leq N$,
 8       and compute departing traffic $desc_{i,j}^{(n)}$, $1 \leq j \leq N$.
 9 **while** $dist(desc^{(n)}, desc^{(n-1)}) > \varepsilon$
10 compute network-wide performance results
```

In each iteration the queueing stations are analyzed using the available descriptions of the traffic arriving at the stations (line 7). The analysis allows to compute station-related performance measures, such as the mean queue length, and, more important, the description of the traffic leaving the stations (line 8). In this way, a new set of traffic descriptors $desc^{(n)} = \{desc_{i,j}^{(n)} | i, j\}$ is computed in each iteration.

When the algorithm starts only the descriptions of the traffic arriving from outside are known (they are part of the model specification). Hence, all other descriptors are initially set to the *null* value in line 2 and have to be ignored in line 7 until a first approximation is available.

The algorithm stops when the distance $dist(desc^{(n-1)}, desc^{(n)})$ between two successive sets of descriptors is smaller than or equal a given threshold ε (line 9). Once all traffic descriptors are known, network-wide performance results can be computed in line 10.

The analysis of individual stations

For the analysis of the single stations (in line 7 and 8 of the iteration algorithm), we define that a station specification consists of two components:

1. a queue with finite or infinite capacity,
2. one or more service entities that serve the jobs (served jobs leave the queue),

and two policies:

1. a policy that handles incoming jobs if the queue is full (only for finite queues),
2. a scheduling policy that describes how the service stations fetch new jobs from the queue.

Such queueing stations can be analyzed by different approaches. Since most analysis methods for queueing processes require that a queueing station has exactly *one* arrival traffic descriptor and *one* traffic departure descriptor, a *traffic merging* (or *traffic superpositioning*) and a *traffic splitting* step are required. The traffic merging step merges for a station its arrival descriptors into a single overall arrival descriptor whereas the traffic splitting step splits the overall departure descriptor into the required number of departure descriptors. Thus, every time when the fixed-point iteration algorithm analyzes a queueing station we have to perform the following steps (as part of steps 7 and 8 in the algorithm above, and illustrated in Figure 15.1 below):

1. merge the incoming traffic;
2. analyze the service operation;
3. split the departure traffic.

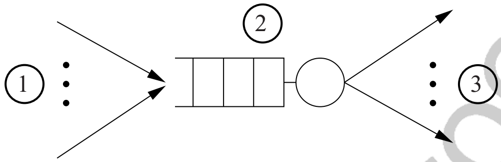


Fig. 15.1: The three key operations to be performed on a traffic stream: merging, queueing and splitting

Notice that for Jackson queueing networks (cf. Section 15.7), these three steps are extremely simple, hence, they are not often distinguished explicitly. Furthermore, for JQNs there appears to be no need for a fixed-point iteration. However, this is only partly true. One can argue that an iterative method to solve the traffic equations in JQNs (like Gauss-Seidel iterations) in fact forms a fixed-point iteration in itself. The distinguishing feature is then that for JQNs, no queueing analysis takes place within the fixed-point computation (only afterwards), whereas, in general, decomposition-based methods do require the intertwining of fixed-point iteration steps and queueing analysis steps, as will become clear in the following sections.

15.3 Whitt's Queueing Network Analyzer

In the early 1980s, Whitt presented the Queueing Network Analyzer (QNA) [49, 50], a software package developed at Bell Laboratories for the approximate analysis of open queueing networks. Unlike prior approaches which were based on Markovian models, QNA allows for the analysis of open queueing networks where the external arrival processes need not be Poissonian and the service times need not be negative exponentially distributed. Additionally, QNA is able to perform the analysis fast: due to the involved approximations and assumptions, the network traffic analysis is, in essence, reduced to the solution of a set of linear equations, comparable to those in JQNs (cf. Section 15.7).

In the following, we will give an overview of the functionality of QNA. The structure of our presentation slightly differs from Whitt's original paper [49], however, it follows the presentation of JQNs in Section 15.7.

15.3.1 Model class

QNA allows for the approximate analysis of open queueing networks fed by external arrival processes, in which the routing takes place according to fixed probabilities (like in JQNs). The nodes are GI|G|m multiserver queues without capacity constraints and with the FCFS service discipline. The external arrival processes as well as the service processes of the nodes are described by the first and the second moment of the inter-arrival, resp. service time distributions. The QNA approach allows for the separate analysis of the nodes, hence, QNA is well scalable to larger networks.

QNA's model class includes three features which we will not describe in detail in the following. First, QNA is able to analyze networks with multiple classes of customers, and secondly, networks with immediate feedback are allowed. Both features are "implemented" by adding a pre-processing and post-processing phase to the core QNA algorithms, that is, QNA treats multiple visits of a single job to one queue as one longer visit, and multiple classes are treated as one class with multimodal service times. The third feature, the customer multiplication factor of a node, only requires small modifications in the service operation equations. Although these features are interesting as such, they have not been implemented for FiFiQueues, however, also in that context they could be added via appropriate pre- and post-processing phases.

15.3.2 Traffic descriptors

The external arrival processes are specified by the first and second moment of the inter-arrival times. In fact, this representation is also applied to the traffic streams between the nodes. More specifically, QNA uses the traffic descriptor $\langle \lambda, c^2 \rangle$ to describe a traffic stream where λ is the arrival rate and c^2 is the squared coefficient of variation of the inter-arrival time.

Clearly, this allows the representation of non-Poissonian processes. However, neither higher moments nor correlations of the arrival stream are considered, which may influence the quality of the analysis. QNA employs fine-tuned heuristics deduced from simulation studies to reduce the errors introduced by this simplification.

15.3.3 Superposition of traffic streams

To merge n traffic streams specified by $\langle \lambda_1, c_1^2 \rangle, \ldots, \langle \lambda_n, c_n^2 \rangle$ into one traffic stream $\langle \lambda, c^2 \rangle$, QNA first computes the total arrival rate which is simply given by

$$\lambda = \sum_{i=1}^{n} \lambda_i.$$

QNA's efficiency is based on the fact that it computes the traffic descriptors from linear systems of equations. The above expression for λ is clearly linear in λ_i. For c^2 a linear equation can be found, too, by the asymptotic approximation method (AS):

$$c_{AS}^2 = \sum_{i=1}^{n} \frac{\lambda_i}{\lambda} c_i^2.$$

However, the asymptotic method does not work well for a wide range of cases. It is therefore combined with the stationary-interval method (SI), resulting in the following hybrid approximation:

$$c^2 = w \cdot c_{AS}^2 + (1-w) \cdot c_{SI}^2.$$

The stationary-interval method does not provide a linear expression for c_{SI}^2, but experiments have shown that setting c_{SI}^2 to 1 (in the expression above) increases the average error only by 1 percent, so that we obtain

$$c^2 = w \cdot c_{AS}^2 + (1-w).$$

Simulations have shown that the above approximations do impact the quality of the analysis of a node which takes the merged traffic stream as input. To improve the results, QNA respects the utilization ρ of the node in the computation of the factor w. With $\rho = \lambda/\mu$ (where μ is the service rate of the queueing station), QNA sets

$$w = \left[1 + 4(1-\rho)^2(v-1)\right]^{-1} \text{with } v = \left(\sum_{i=1}^{n} \left(\frac{\lambda_i}{\lambda}\right)^2\right)^{-1}.$$

15.3.4 Splitting traffic streams

When splitting, QNA assumes that the involved processes are renewal processes. Under this assumption, an exact solution is available. For n splitting probabilities $p_1, \ldots, p_n$ and the traffic stream $\langle \lambda, c^2 \rangle$, we obtain the splitted streams

$$\langle \lambda_1, c_1^2 \rangle, \ldots, \langle \lambda_n, c_n^2 \rangle,$$

with

$$\lambda_i = p_i \cdot \lambda, \quad \text{and} \quad c_i^2 = p_i \cdot c^2 + (1-p_i), \qquad i = 1, \ldots, n.$$

15.3.5 Servicing jobs

Network nodes are analyzed as GI|G|m queues. Let $\langle \lambda_A, c_A^2 \rangle$ be the arrival traffic descriptor of the node and m the number of service entities. The service process is specified by the service rate μ and by the squared coefficient of variation c_S^2 of the service time distribution. We require the stability of all stations, i.e., $\lambda_A < \mu$. How does QNA compute the departure descriptor $\langle \lambda_D, c_D^2 \rangle$?

Since the queues are stable and have infinite capacity, no losses occur and we clearly have $\lambda_D = \lambda_A$. To compute c_D^2, Whitt combines Marshall's formula [33] with other approximations to obtain

$$c_D^2 = 1 + (1-\rho^2)(c_A^2 - 1) + \frac{\rho^2}{\sqrt{m}}(c_S^2 - 1). \qquad (15.1)$$

The involved approximations may lead to large errors when c_S^2 is small, thus QNA uses the following extension of the above formula:

$$c_D^2 = 1 + (1-\rho^2)(c_A^2 - 1) + \frac{\rho^2}{\sqrt{m}}(\max\{c_S^2, 0.2\} - 1). \qquad (15.2)$$

Note again the linearity of the expressions for λ_D and c_D^2 in the arrival traffic $\langle \lambda_A, c_A^2 \rangle$.

15.3.6 Node performance

QNA is able to compute results for the first and second moment of the waiting time W and the queue length N. Due to the complexity of the involved approximations, we limit our presentation only to the simplest one, i.e., the computation of $\mathrm{E}[W]$ in the case of single-server GI|G|1 queues. The required derivations for the other quantities can be found in [49, Eq. (46)–(71)]. For given arrival traffic $\langle \lambda_A, c_A^2 \rangle$, service descriptor $\langle \mu, c_S^2 \rangle$ and utilization ρ, $\mathrm{E}[W]$ is approximated as

$$\mathrm{E}[W] = \frac{\rho}{2(1-\rho)\mu}(c_A^2 + c_S^2) g(\rho, c_A^2, c_S^2), \qquad (15.3)$$

where the function ρ is defined as

$$g(\rho, c_A^2, c_S^2) = \begin{cases} \exp\left(\frac{2(1-\rho)(1-c_A^2)^2}{3\rho(c_A^2+c_S^2)}\right), & c_A^2 < 1, \\ 1, & c_A^2 \geq 1. \end{cases}$$

Note that Equation (15.3) is exact for $c_A^2 = 1$, i.e., in the case of an M|G|1 queue. When $c_A^2 < 1$, it is equivalent to the Krämer and Langenbach-Belz approximation [26].

15.3.7 Network-wide performance

The results presented for network performance measures in Jackson queueing networks (see Section 15.7) can also be applied here, providing expressions for $\mathrm{E}[V_i]$, $\mathrm{E}[T_i]$, $\mathrm{E}[T_{total}]$ and $\mathrm{E}[N_{total}]$. Additionally, Whitt developed approximations for the variances of the above-stated measures [49, Eq. (80)–(84)].

15.3.8 Complexity

In the above sections, we have repeatedly pointed out the linearity of the employed equations for the three traffic operations merging, splitting, and service. In fact, QNA exploits this linearity to efficiently evaluate the queueing network.

First, for the arrival rates of the traffic streams the system of equations derived for JQNs is also valid for QNA. Let $\left\langle \lambda_{A,i}, c^2_{A,i} \right\rangle$ be the traffic arriving at node i, $\left\langle \lambda_{D,i}, c^2_{D,i} \right\rangle$ the traffic leaving this node, and $\left\langle \lambda_{ext,i}, c^2_{ext,i} \right\rangle$ the external traffic. If $\Gamma = (r_{ij})$ is the routing matrix, the following traffic equation holds for each node $i = 1, \ldots, n$ of the network:

$$\lambda_{A,i} = \lambda_{ext,i} + \sum_{j=1}^{n} \lambda_{D,j} \cdot r_{ji}. \tag{15.4}$$

Again, QNA's model class implies $\lambda_{D,i} = \lambda_{A,i}$ and the traffic equations form a system of linear equations which can be expressed in vector/matrix notation as

$$\lambda_{\mathrm{A}} = \lambda_{\mathrm{ext}}(\mathrm{I} - \Gamma)^{-1}.$$

For the squared coefficients of variation of the traffic streams a system of equations can be set up, too. The synthesis of the superposition and the splitting operations yields

$$c^2_{A,i} = (1 - w_i) + w_i \left(p_{ext,j} c^2_{ext,i} + \sum_{j=1}^{n} p_{j,i} (r_{ji} c^2_{D,j} + 1 - r_{ji}) \right),$$

where $p_{j,i} = \lambda_{D,j} r_{j,i} / \lambda_{A,i}$ is the fraction of traffic arriving from node j to node i and $p_{ext,j} = \lambda_{ext,i} / \lambda_{A,i}$ is the fraction of external traffic arriving to node i. Finally, if we include the result of the service operation we obtain the following system of linear equations

$$\begin{aligned} c^2_{A,i} = (1 - w_i) + w_i \{ p_{ext,j} c^2_{ext,i} \; + \; & \sum_{j=1}^{n} p_{j,i} (r_{ji} (1 + (1 - \rho_i^2)(c^2_{A,j} - 1) \\ & + \frac{\rho_i^2}{\sqrt{m_i}} (\max(c^2_{S,i}, 0.2) - 1)) + 1 - r_{ji}) \}. \end{aligned} \tag{15.5}$$

Using the equations (15.4–15.5), the traffic descriptors can easily be computed. Thus, obviously QNA has the same time complexity as the Jackson network method. Note that, due to the linearity of the involved equations, QNA does not require the fixed-point iteration described in Section 15.2 (although, if an iterative solver is used to solve the linear equations, the fixed-point iteration can be regarded as hidden in the solver).

15.4 FiFiQueues

In the mid-1990's Haverkort and Weerstra, cf. [13, 14, 15, 48], extended Whitt's QNA approach by means of replacing the core of the analysis: the service operation. Unlike QNA, their new approach, called QNAUT, does not directly use the descriptor of the arrival traffic to compute the departure traffic descriptor, but assumes that the arrival traffic descriptor can be used to construct a phase-type (PH) renewal process (see Section 15.8.2) which approximates the "real" underlying arrival process. This allows for the inclusion of finite-buffer queueing stations as well as for the analysis of the queueing stations by matrix-geometric and general Markovian techniques, instead of the approximations originally used in QNA.

At the end of the 1990s, an extended version of the original approach was proposed, in which some approximate steps were removed and the model class was slightly enhanced [41, 42, 43]. In particular, this enhanced class provides:

- exact results for the the departure process based on the results of Bocharov [5] for PH$|$PH$|1|K$ queues;
- efficient per-queue analysis;
- for each finite queueing station, a traffic stream is computed which consists of the customers rejected at a completely filled queue. This loss traffic stream can be used as arrival stream for other queueing stations like any other "regular" departure traffic stream.

This approach, as well as the analysis tool developed from it, is named *FiFiQueues* (for *Fi*xpoint-based analysis of networks with *Fi*nite *Queues*).

15.4.1 Model class

The external arrival processes are described, as in QNA, by the first and the second moment of the inter-arrival times. The main differences to QNA's model class are:

- the service processes are specified as PH renewal processes;
- the queueing stations can have *infinite* or *finite* queueing capacities. The nodes are analyzed as PH$|$PH$|1(|K)$ queues with the FCFS service discipline. The customer multiplication factor known from QNA is also supported, but not described in the following;

- finite queues have two output streams: the "regular" departure traffic stream and the loss traffic stream which consists of the customers rejected by a full queue.

Seen from a single queue, customers arriving at a completely filled queue are simply lost. This form of blocking is common in communication networks (communication blocking) and has an important advantage: unlike other types of blocking (like back-blocking), it still allows the independent analysis of each of the queueing stations.

Just like the regular departure traffic of a queueing station with finite capacity, the loss traffic is not known a priori and is computed by the analysis of the station. The "reuse" of loss traffic streams as arrival streams to other nodes requires an auxiliary routing matrix. Its handling will not be discussed further in the following sections, since, once the traffic descriptors of the loss streams are known, they can easily be processed like the regular departure traffic. However, note that loss traffic streams should only be used very carefully in feedback networks: if a loss traffic stream is fed back directly or indirectly to the node which produced the stream, it can prevent the iteration algorithm (see Section 15.2) to terminate because the arrival rate to the node increases in each iteration step.

15.4.2 Traffic descriptor

As in QNA, the external arrival processes as well as the inter-node traffic streams are described by the first and second moment of the inter-arrival times. The traffic descriptor $\langle \lambda, c^2 \rangle$ contains the arrival rate λ and the squared coefficient of variation c^2 of the inter-arrival time.

15.4.3 Superposition of traffic streams

To merge n traffic streams specified by $\langle \lambda_1, c_1^2 \rangle, \ldots, \langle \lambda_n, c_n^2 \rangle$ into one traffic stream $\langle \lambda, c^2 \rangle$, we adopt the hybrid approximation of QNA, i.e.,

$$\lambda = \sum_{i=1}^{n} \lambda_i, \tag{15.1}$$

$$c^2 = w \cdot \sum_{i=1}^{n} \frac{\lambda_i}{\lambda} c_i^2 + (1-w), \tag{15.2}$$

with

$$w = \left[1 + 4(1-\rho)^2(v-1)\right]^{-1}, \quad \text{and} \quad v = \left(\sum_{i=1}^{n} \left(\frac{\lambda_i}{\lambda} \right)^2 \right)^{-1},$$

where ρ is the utilization of the node receiving the resulting traffic stream. It should be emphasized that these formulae were originally designed in the context of QNA's model class, i.e., *not* for finite queues. Thus, their usage in FiFiQueues introduces auxiliary errors to the computation, in addition to the errors inherent to the hybrid approximation method.

One may wonder if we could obtain better results by not following QNA's linear approximation ($c_{SI}^2 = 1$) but in actually computing the correct value for c_{SI}^2. Our experiments have shown that nearly the same results are obtained by doing so. This is consistent with Whitt's observation that fixing c_{SI}^2 at 1 increases the average error by only 1 percent.

15.4.4 *Splitting traffic streams*

When splitting, we assume that the involved processes are renewal processes. Under this assumption, an exact solution is available. For n splitting probabilities $p_1, \ldots, p_n$ and the traffic stream $\langle \lambda, c^2 \rangle$, we obtain the splitted streams $\langle \lambda_1, c_1^2 \rangle, \ldots, \langle \lambda_n, c_n^2 \rangle$ with

$$\lambda_i = p_i \cdot \lambda, \quad \text{and} \quad c_i^2 = p_i \cdot c^2 + (1 - p_i), \qquad i = 1, \ldots, n. \tag{15.3}$$

15.4.5 *Servicing jobs*

We have already stated that the nodes are analyzed as PH|PH|1(|K) queues. Thus, before a queueing station can be analyzed we need to find a PH distribution that fits the two moments given in the arrival traffic descriptor. In the following we will explain the fitting step and the actual queueing analysis procedure, thereby treating PH|PH|1 and PH|PH|1|K queues separately. We require that the PH|PH|1 queues are stable, i.e., the total arrival rate at a PH|PH|1 station should be smaller than its service rate.

15.4.5.1 Phase-type representation of the arrival processes

Let $\langle \lambda, c^2 \rangle$ be the arrival traffic descriptor. We write $\mathrm{E}[X] = 1/\lambda$ for the corresponding mean inter-arrival time. Clearly, having only two moments allows us some freedom to select an appropriate PH distribution. We require that the chosen PH distribution, represented by (α, A)

1. matches the two moments exactly (at least for a certain range; see below), and
2. is as compact as possible, i.e., has the smallest number of transient states m.

Additionally, we want that the employed fitting procedure does not consume too much time since it has to be executed every time when a node is analyzed. In FiFiQueues, we use the following approach, first presented in [14]. Two cases are distinguished:

- In case $c^2 \leq 1$, we use a hypo-exponential distribution with $m = \left\lceil \frac{1}{c^2} \right\rceil$ phases and initial probability vector $\alpha = (1, 0, \cdots, 0)$. The matrix A is then given as

$$\mathrm{A} = \begin{pmatrix} -\lambda_0 & \lambda_0 & & & \\ & -\lambda_1 & \lambda_1 & & \\ & & \ddots & \ddots & \\ & & & -\lambda_{m-2} & \lambda_{m-2} \\ & & & & -\lambda_{m-1} \end{pmatrix}, \qquad (15.4)$$

where $\lambda_i = m/\mathrm{E}[X]$, for $0 \leq i < m-2$ and where

$$\lambda_{m-1} = \frac{2m\left(1 + \sqrt{\frac{1}{2}m(mc^2 - 1)}\right)}{\mathrm{E}[X](m + 2 - m^2c^2)} \text{ and } \lambda_{m-2} = \frac{m\lambda_{m-1}}{2\lambda_{m-1}\mathrm{E}[X] - m} .$$

For small c^2, PH distributions with a large number of states will be obtained. To limit the computational requirements in the analysis process we do not allow c^2 to be smaller than $\frac{1}{10}$. This approximation corresponds to an Erlang-10 distribution and produces generally good results, also as approximation for deterministic distributions.

- In case $c^2 > 1$, we take a hyper-exponential distribution with $m = 2$ phases. Such a distribution has three free parameters: the choice probability p between the two possible phases and the rates μ_1 and μ_2 of the two phases. Fitting the first two moments thus leaves one degree of freedom. We resolve this by assuming so-called "balanced means", meaning that the ratios p/μ_1 and $(1-p)/\mu_2$ should be equal. This then yields $\alpha = (p, 1-p)$ and

$$\mathrm{A} = \begin{pmatrix} -\frac{2p}{\mathrm{E}[X]} & 0 \\ 0 & -\frac{2(1-p)}{\mathrm{E}[X]} \end{pmatrix} \text{ with } p = \frac{1}{2} + \frac{1}{2}\sqrt{\frac{c^2 - 1}{c^2 + 1}} .$$

15.4.5.2 Analysis of PH|PH|1|K queues

The underlying CTMC Let (α, A) be the arrival PH renewal process with l states as obtained by the fitting step and (β, B) the service PH renewal process with m states. Then we can describe the behavior of a node with queueing capacity K by a QBD process [37] (see Section 15.8.4) with $K+1$ levels, where level 0 consists of l states and where levels 1 through K consist of $l \cdot m$ states each.

The i-th level represents the state of the system when it contains i customers. A step from level i to level $i+1$ ($i < K$) stands for an arrival and a step from level i to level $i-1$ ($i > 0$) stands for a departure. The $l \cdot m$ states of a level $i > 0$ describe the current state of the arrival and of the service processes (level 0 contains only l states because the queue is empty and the service process has not yet started; it only records the state of the arrival process). This leads to the following generator matrix of the Markov chain:

$$\mathrm{Q} = \begin{pmatrix} \mathrm{A} & \mathrm{A}^0\alpha \otimes \beta & & \\ \mathrm{I} \otimes \mathrm{B}^0 & \mathrm{A} \oplus \mathrm{B} & \mathrm{A}^0\alpha \otimes \mathrm{I} & 0 \\ 0 & \mathrm{I} \otimes \mathrm{B}^0\beta & \mathrm{A} \oplus \mathrm{B} & \mathrm{A}^0\alpha \otimes \mathrm{I} \\ & \ddots & \ddots & \ddots \\ & & \mathrm{I} \otimes \mathrm{B}^0\beta & (\mathrm{A} + \mathrm{A}^0\alpha) \oplus \mathrm{B} \end{pmatrix},$$

where $\mathrm{A}^0 = -\mathrm{A} \cdot 1$, $\mathrm{B}^0 = -\mathrm{B} \cdot 1$, $\mathrm{L} \oplus \mathrm{M} = \mathrm{L} \otimes \mathrm{I} + \mathrm{I} \otimes \mathrm{M}$, and $\otimes$ is the Kronecker product operator (also known as tensor or matrix direct product operator).

The steady-state solution v of the Markov chain with generator Q can be obtained by solving the global balance equation (see Section 15.8.4):

$$\mathrm{v} \cdot \mathrm{Q} = 0 \quad \text{and} \quad \mathrm{v} \cdot 1 = 1.$$

The vector v is of size $l + K \cdot l \cdot m$. In the following we write v_0 for the vector $(v_1, \ldots, v_l)$ which contains the steady-state probabilities of level 0 and we write v_i for the vector $(v_{l+1+(i-1)\cdot l \cdot m}, \ldots, v_{l+i \cdot l \cdot m})$ which contains the steady-state probabilities of level $i = 1, \ldots, K$.

The departure traffic The steady-state solution vector v now allows us to compute the departure traffic descriptor $\langle \lambda_D, c_D^2 \rangle$. To this end, we use the results of Bocharov presented in [5] which we will briefly describe in the following.

We begin with the computation of the blocking probability π, i.e., the probability that an arriving customer encounters a full queue and, hence, is lost. The vector $\mathrm{v}_{\mathrm{A,K}}$ gives for this situation the state probabilities and it holds that

$$\mathrm{v}_{\mathrm{A,K}} = \frac{1}{\lambda_A} \mathrm{v}_\mathrm{K} (\mathrm{A}^0 \otimes \mathrm{I}),$$

where λ_A stands for the arrival rate to the node and K stands for the queueing capacity of the node. This leads to the blocking probability π:

$$\pi = \mathrm{v}_{\mathrm{A,K}} \cdot 1.$$

With π, we easily find the departure rate of served customers as

$$\lambda_D = \lambda_A (1 - \pi). \tag{15.5}$$

Higher moments of the inter-departure time can be computed using the following consideration. If the queue is not empty after a departure took place, the distribution of the time up to the next departure is equal to the distribution of the service time. Otherwise, it is equal to the distribution of the sum of the time until the next customer arrival and its service time (which are independent). The probability to leave an empty queue at departure instant $t + \varepsilon$ is

$$\mathrm{v}_{\mathrm{D},0} = \frac{1}{\lambda_D}\mathrm{v}_1(\mathrm{I}\otimes\mathrm{B}^0). \tag{15.6}$$

This leads Bocharov to the expression for the i-th moment d_i of the inter-departure time distribution:

$$d_i = b_i + \mathrm{v}_{\mathrm{D},0}\sum_{j=1}^{i}(-1)^j\frac{i!}{(i-j)!}\mathrm{A}^{-j}1b_{i-j}, \tag{15.7}$$

where b_i is the i-th moment of the service time distribution. Thus, one can easily verify that the variance σ_D^2 of the departure process is

$$\sigma_D^2 = \sigma_S^2 + \sigma_0^2, \tag{15.8}$$

where σ_S^2 is the variance of the service time distribution and σ_0^2 equals

$$\sigma_0^2 = 2\mathrm{v}_{\mathrm{D},0}\mathrm{A}^{-2}1 - (\mathrm{v}_{\mathrm{D},0}\mathrm{A}^{-1}1)^2. \tag{15.9}$$

The squared coefficient of variation is then given by $c_D^2 = \lambda_D^2\sigma_D^2$.

The loss traffic The rate of loss λ_L is given by $\lambda_L = \lambda_A\cdot\pi$, where π is the loss probability. In oder to obtain higher moments of the inter-loss time we describe the loss process by the MAP $(\mathrm{L}_0,\mathrm{L}_1)$ with

$$\mathrm{L}_0 = \begin{pmatrix} \mathrm{A} & \mathrm{A}^0\alpha\otimes\beta & & & \\ \mathrm{I}\otimes\mathrm{B}^0 & \mathrm{A}\oplus\mathrm{B} & \mathrm{A}^0\alpha\otimes\mathrm{I} & & \\ 0 & \mathrm{I}\otimes\mathrm{B}^0\beta & \mathrm{A}\oplus\mathrm{B} & \mathrm{A}^0\alpha\otimes\mathrm{I} & \\ & \ddots & \ddots & \ddots & \\ & & \mathrm{I}\otimes\mathrm{B}^0\beta & \mathrm{A}\oplus\mathrm{B} & \end{pmatrix}, \mathrm{L}_1 = \begin{pmatrix} 0 & & \\ & \ddots & \\ & & \mathrm{A}^0\alpha\otimes\mathrm{I} \end{pmatrix}.$$

The underlying CTMC of this MAP is the CTMC of the QBD where arrivals in the last level K have been marked. Naturally, it has the same steady-state probability vector v. The i-th moment of the inter-loss time is given by

$$\mathrm{E}[L^i] = \frac{i!}{\lambda_D}\mathrm{v}(-\mathrm{L}_0)^{-(i-1)}1, \tag{15.10}$$

hence, its second moment equals

$$\mathrm{E}[L^2] = \frac{2}{\lambda_L}\mathrm{v}(-\mathrm{L}_0)^{-1}1.$$

15.4.5.3 Analysis of PH|PH|1 queues

The underlying CTMC Let (α,A) be the arrival PH renewal process with l states and (β,B) the service PH renewal process with m states. Again, the behavior of the queue can be described by a QBD process with a generator matrix similar to the one of the PH|PH|1|K; the only difference is the fact that it has repeating columns ad infinitum:

$$\mathrm{Q} = \begin{pmatrix} \mathrm{A} & \mathrm{A}^0\alpha\otimes\beta & & \\ \mathrm{I}\otimes\mathrm{B}^0 & \mathrm{A}\oplus\mathrm{B} & \mathrm{A}^0\alpha\otimes\mathrm{I} & \\ 0 & \mathrm{I}\otimes\mathrm{B}^0\beta & \mathrm{A}\oplus\mathrm{B} & \mathrm{A}^0\alpha\otimes\mathrm{I} \\ & \ddots & \ddots & \ddots \end{pmatrix},$$

with the infinite steady-state probability vector v fulfilling

$$\mathrm{v}\cdot\mathrm{Q} = 0 \quad \text{and} \quad \mathrm{v}\cdot 1 = 1.$$

We refer to Section 15.8.3 for an overview of solution techniques.

The departure traffic Since infinite queues produce no loss, we have

$$\lambda_D = \lambda_A, \tag{15.11}$$

where λ_A is the arrival rate to the node. The variance of the output stream is calculated using the same approach as in the case of finite-buffer queues and the equations (15.6), (15.8), and (15.9) still hold.

15.4.6 Node performance

FiFiQueues computes the first and second moment of the waiting time W and the queue length N. Again, queues with finite and infinite buffer capacity are treated separately.

15.4.6.1 Node performance of PH|PH|1|K queues

The j-moment $\mathrm{E}\left[N^j\right]$ of the queue length distribution (including the job in service) is given by

$$\mathrm{E}\left[N^j\right] = \sum_{i=1}^{K} i^j \mathrm{v_i} 1. \tag{15.12}$$

Hence, mean and variance of the queue length N are:

$$\mathrm{E}\left[N\right] = \sum_{i=1}^{K} i\cdot\mathrm{v_i}1 \quad \text{and } \mathrm{Var}\left[N\right] = \sum_{i=1}^{K} i^2\cdot\mathrm{v_i}1 - \mathrm{E}\left[N\right]^2.$$

Equation (4.4) in [5] gives the Laplace-Stieltjes transform of the waiting time probability density function. From this equation, any desired moment of the waiting time can be derived. For the mean and the variance we obtain [5, Eq. (4.5)–(4.7)]:

$$
\begin{aligned}
\mathrm{E}[W] &= \frac{1}{\lambda_D}(\mathrm{E}[N] - 1 + \mathrm{v}_0 1),\\
\mathrm{Var}[W] &= \frac{2}{\lambda_D}\left(\mu \cdot \mathrm{q}_2 1 - \left(\mathrm{q}_1\left(1 \otimes \mathrm{B}^{-1} 1\right)\right)\right) - \mathrm{E}[W]^2,
\end{aligned}
$$

where μ is the service rate. The components of the vector q_1 resp. q_2 give the first, resp. second binomial moment of the number of jobs in the queue as a function of the system state. For $j > 0$, the j-th binomial moment q_j is defined as [5, Eq. (3.1)]:

$$
\mathrm{q}_j = \sum_{i=j+1}^{K} \binom{i-1}{j} \mathrm{v}_\mathrm{i}.
$$

15.4.6.2 Node performance of PH|PH|1 queues

In the case of infinite buffer capacity, the expressions presented for the PH|PH|1|K queue in the previous section can still be applied, provided that the steady-state probability vectors v_i are available in a form that allows to calculate the, now infinite, sums. For example, if we assume that a matrix-geometric solution method (see Section 15.8.3) is employed to compute the steady-state probabilities, the vectors v_i have the so-called matrix geometric form

$$
\mathrm{v}_\mathrm{i} = \mathrm{v}_1 \mathrm{R}^{i-1}, \quad \mathrm{R} \in \mathbb{R}^{lm \times lm}, \quad i = 1, 2, \ldots,
$$

where R is the entry-wise smallest non-negative solution of the matrix-quadratic equation

$$
\mathrm{A}^0 \alpha \otimes \beta + \mathrm{R}(\mathrm{A} \oplus \mathrm{B}) + \mathrm{R}^2(\mathrm{I} \otimes \mathrm{B}^0 \beta) = 0.
$$

The j-th moment of the queue length distribution is then given by

$$
\mathrm{E}[N^j] = \sum_{i=1}^{\infty} i^j \mathrm{v}_\mathrm{i} 1 = \sum_{i=1}^{\infty} i^j \mathrm{v}_1 \mathrm{R}^{i-1} 1, \tag{15.13}
$$

which yields in case $j = 1$:

$$
\mathrm{E}[N] = \mathrm{v}_1 (\mathrm{I} - \mathrm{R})^{-2} 1.
$$

Similarly, the other node performance measures can be obtained.

15.4.7 Network-wide performance

Many results for the network performance measures developed by Whitt for QNA (see Section 15.3.7) can also be applied to FiFiQueues when respecting the fact that, due to losses at finite queues, the departure rate of a node may differ from the total arrival rate to that node. Additionally, one has to decide how loss traffic streams should be treated in the computation of network-wide performance results. For example, the following question has to be answered: should the expected number of visits $\mathrm{E}[V_i]$ also include rejections due to full buffers? As this is only a problem of "interpretation" of the results, we will not discuss it further here.

15.4.8 Complexity

15.4.8.1 Traffic computation

In FiFiQueues the traffic descriptor of the outgoing traffic depends in a complex, non-linear way on the incoming traffic. Thus, unlike the QNA method, FiFiQueues clearly requires an iterative computation scheme to compute the descriptors of the internal traffic streams. A deeper discussion of FiFiQueues' iteration behavior is given in Section 15.5. Here, we will analyze the complexity of the operations that have to be performed for each node during each iteration.

First, we can safely neglect the traffic merging and splitting steps in our discussion. They only consist of a small number of additions and multiplications. The most time and space consuming operation is the service operation. It can be divided into three phases:

1. fitting of the PH distribution to the arrival traffic,
2. computation of the steady-state probability vector of the underlying CTMC, and
3. computation of the departure traffic descriptor (and, if needed, of the loss traffic descriptor).

Again, we can neglect the first phase since its time complexity is $O(1)$. For the second phase, we distinguish between finite and infinite queueing stations.

If the queueing capacity is finite, so is the CTMC. Let l be the size of the arrival PH process, i.e., the number of states of its CTMC representation, m the size of the service PH process and K the queueing capacity. Then, the generator matrix is of size $(l+lmK)\times(l+lmK)$. This corresponds to a finite QBD with $N_0 = l$ and $N = lm$ (see Section 15.8.4). The latest implementation of FiFiQueues uses for finite capacities the Cyclic Reduction method [3] which has time complexity $O((l+m)^3\log K+(l+m)^2K)$. If the descriptor of the loss traffic is required, additional operations have to be performed to compute the product $\mathrm{v}(-\mathrm{L}_0)^{-1}$. For unbounded queueing capacity, the LR algorithm [28] is used.

Once the steady-state solution is known, the departure traffic descriptor can be computed. Both for finite and infinite queueing stations, this only requires a small number of matrix vector multiplications. Note that the moments b_i of the service process needed by Equation (15.7) are constant for a given network and hence can be precomputed once.

15.4.8.2 Node performance and network performance computation

Since the network performance computation is comparable to that of the QNA method, we only discuss the complexity of the node performance computation here.

Concerning finite queueing stations, the computation of the mean and variance of the queue length requires the summation over the $lm(K-1)$ entries of the steady-state probability vector. For the moments of the waiting time distribution, we have to invert matrix B of size $m \times m$ which can be seen as a constant time operation even for very complex PH representations of the service process (say, $m = 50$).

In case of infinite queueing capacity, the complexity depends on the employed solution method. Assuming a matrix-geometric solution method, the expression $\mathrm{E}[N] = \mathrm{v}_1(\mathrm{I}-\mathrm{R})^{-2}1$ we gave for the mean queue length in Section 15.4.6, requires the vector $\mathrm{X} = \mathrm{v}_1(\mathrm{I}-\mathrm{R})^{-2}$ which can be obtained by solving the linear system $\mathrm{X}(\mathrm{I}-\mathrm{R})^2 = \mathrm{v}_1$ of order lm.

15.4.9 The FiFiQueues network designer

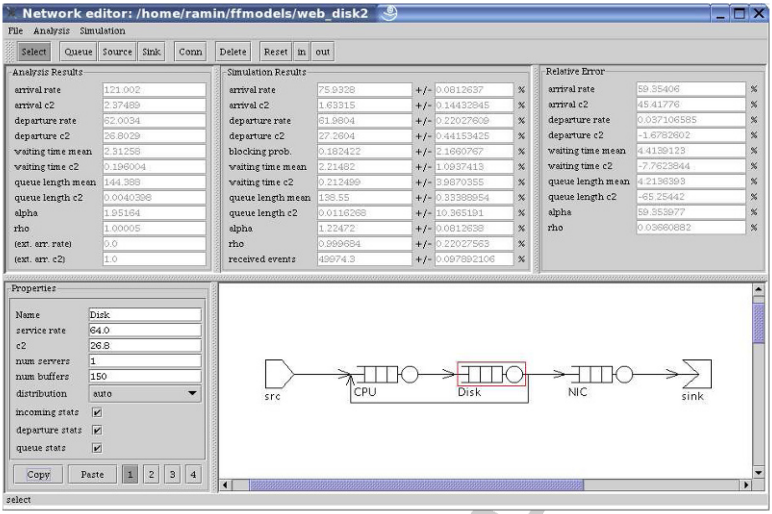
The FiFiQueues approach has proven to be stable and reliable enough for end users. In this section we present an integrated tool environment, the FiFiQueues network designer, that allows an easy access to the underlying algorithms. The tool also contains a simulator for the steady-state simulation of queueing networks. The FiFiQueues network designer consists of a graphical user interface written in Java, a numerical analysis module, and a simulation module. The latter two have been written in C++.

15.4.9.1 The graphical user interface

The graphical user interface allows to construct, edit and study open and closed queueing networks of arbitrary topology. The networks can be evaluated by numerical analysis or by simulation. Figure 15.2 shows a screenshot of the main window. The lower part of the window shows the edited network and the properties of the currently selected node. The upper part displays the results of the numerical analysis (left section) and the results of the simulation (middle section, including the 95% confidence intervals) as well as a comparison of both methods (right section).

Every object in the network has properties that can be edited via the user interface. Figure 15.3(a) shows the properties of a finite queueing station while the user is selecting a service distribution. The global-properties panel (see Figure 15.3(b)) allows to control the length of the simulation and the parameters specific to closed networks.

The user interface communicates with the numerical analyzer and the simulator via text files. As an example, the network shown in the screenshot is translated into the following textual description in order to evaluate it.



Analysis Results

arrival rate	121.002
arrival c2	2.37489
departure rate	62.0034
departure c2	26.8029
waiting time mean	2.31258
waiting time c2	0.196004
queue length mean	144.388
queue length c2	0.0040398
alpha	1.95164
rho	1.00005
(ext. arr. rate)	0.0
(ext. arr. c2)	1.0

Simulation Results

arrival rate	75.9328	+/-	0.0812637	%
arrival c2	1.63315	+/-	0.14432845	%
departure rate	61.9804	+/-	0.22027609	%
departure c2	27.2604	+/-	0.44153425	%
blocking prob.	0.182422	+/-	2.1660767	%
waiting time mean	2.21482	+/-	1.0937413	%
waiting time c2	0.212499	+/-	3.9870355	%
queue length mean	138.55	+/-	0.33388954	%
queue length c2	0.0116268	+/-	10.365191	%
alpha	1.22472	+/-	0.0812638	%
rho	0.999684	+/-	0.22027563	%
received events	49974.3	+/-	0.097892106	%

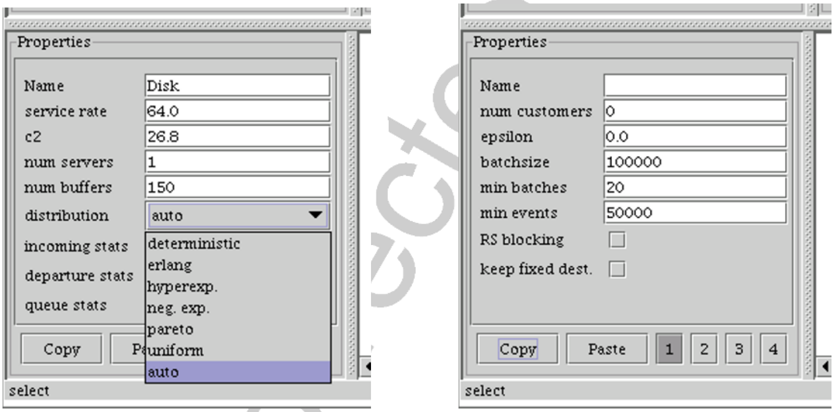
Relative Error

arrival rate	59.35406	%
arrival c2	45.41776	%
departure rate	0.037106585	%
departure c2	-1.6782602	%
waiting time mean	4.4139123	%
waiting time c2	-7.7623844	%
queue length mean	4.2136393	%
queue length c2	-65.25442	%
alpha	59.353977	%
rho	0.03660882	%

Properties

Name	Disk
service rate	64.0
c2	26.8
num servers	1
num buffers	150
distribution	auto
incoming stats	☑
departure stats	☑
queue stats	☑

Fig. 15.2: Main window of the graphical user interface



Properties

Name	Disk
service rate	64.0
c2	26.8
num servers	1
num buffers	150
distribution	auto
incoming stats	
departure stats	
queue stats	

(a) Properties of a network node

Properties

Name	
num customers	0
epsilon	0.0
batchsize	100000
min batches	20
min events	50000
RS blocking	□
keep fixed dest.	□

(b) Global properties of the network

Fig. 15.3: Property editor

```
# Queue mapping
# 0 CPU
# 1 NIC
# 2 Disk
network_props
1 3 1 0 100000 20 50000 0   0   0.0
source_props
90.0 1.2 0 6
```

```
queue_props
1200.0 26.8 1 150 6 1  1  1
1401.0 26.8 1 150 6 1  1  1
64.0 26.8 1 150 6 1  1  1
counter_dest
-1
r
0.0 0.0 1.0 0.0
0.0 0.0 0.0 1.0
0.5 0.5 0.0 0.0
b
0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0
0.0 0.0 0.0 0.0
```

15.4.9.2 The numerical analysis module

The numerical analysis module is the core of the implementation. It incorporates the FiFiQueues algorithms as discussed and the extension for closed queueing networks as described in [44].

15.4.9.3 The simulation module

The simulation module offers the discrete-event simulation of open and closed queueing networks. It is described in detail in [40].

15.5 Performance of FiFiQueues

In this section we evaluate the performance of the FiFiQueues algorithm with regard to the quality of the numerical results. This evaluation consists of

- tests with the FiFiQueues algorithm on some representative queueing networks (Section 15.5.1),
- a case study of a web server (Section 15.5.2).

The results of the numerical analysis are compared to results determined using discrete-event simulation. The relative half-width of the 95%-confidence intervals is smaller than 1% for all the simulation results. If not stated otherwise, arrival time and service time distributions specified by their rate and the squared coefficient of variation (SCV) are always mapped to PH distributions. Relative errors between numerical analysis and simulation are always computed relative to the latter. We conclude with Section 15.5.3.

15.5.1 Evaluation of FiFiQueues

In this section we evaluate FiFQueues' performance with some typical networks. We begin with a single queue in Section 15.5.1.1, then continue with some complex networks in Section 15.5.1.2. The presented tests cover a wide range of input parameters, including (nearly) deterministic processes, and complex networks with finite queueing capacities.

15.5.1.1 Single queues

In the case of queueing networks that consist of only one queueing station, FiFiQueues always produces exact results, provided that the selected arrival and service PH renewal processes match the actual arrival and service processes of the real system. Hence, results of single-queue systems are not very interesting. At this place, we will only discuss the special case of deterministic distributions.

As explained, FiFiQueues limits the number of phases in hypo-exponential PH distributions to 10, which corresponds to a minimum SCV of 0.1. As a consequence, deterministic distributions can only be approximated. To evaluate the effect of this restriction we have analyzed a queueing station with negative exponential services and deterministic arrival process at different loads. Table 15.1 compares the thus obtained mean queue lengths with results found by simulation. It shows that the relative error between analysis and simulation increases with the load. Errors of comparable magnitude can also be observed for other performance measures and for hypo-exponential and hyper-exponential service distributions.

load	analysis	simulation	rel. error
0.1	0.10	0.10	0.0%
0.2	0.20	0.20	0.0%
0.4	0.47	0.45	4.4%
0.6	0.95	0.89	6.7%
0.8	2.34	2.18	7.3%
0.95	10.60	9.26	14.4%

Table 15.1: Mean queue length for a queueing station with deterministic arrival traffic

15.5.1.2 Queueing networks with feedback

3-node queueing network We first address three queueing nodes in series, with a feedback from the last to the first queue, as shown in Figure 15.4. The external Poisson source has rate 1.3 and the service times are Erlang-5 distributed with rate

1.5; the node capacity is 10 (not including the service station) at all queues. The feedback probability is 25%.

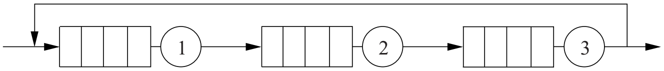


Fig. 15.4: 3-node queueing network with feedback

The results are shown in Table 15.2. The first two rows show the characteristics of the traffic leaving the queueing network from node 3. The middle six rows show the rate and SCV of the arrival traffic at each node, and the last three rows show the expected queue length at each node.

		analysis	simulation	rel. error
Output traffic	$\lambda_{netd,3}$	1.08	1.08	0.0%
	$c^2_{netd,3}$	0.41	0.41	0.0%
Arrival traffic	$\lambda_{a,1}$	1.65	1.66	-0.6%
	$c^2_{a,1}$	0.96	0.96	0.0%
	$\lambda_{a,2}$	1.47	1.47	0.0%
	$c^2_{a,2}$	0.23	0.23	0.0%
	$\lambda_{a,3}$	1.45	1.45	0.0%
	$c^2_{a,3}$	0.21	0.22	-4.5%
Queue length	$\mathrm{E}[N_1]$	6.47	6.47	0.0%
	$\mathrm{E}[N_2]$	4.43	4.45	-0.4%
	$\mathrm{E}[N_3]$	3.96	3.90	1.5%

Table 15.2: Results for the 3-node network with Poisson source

The good results of the analysis can be explained by the fact that the resulting arrival traffic to node 1 (i.e., where the traffic superposition operation happens) is near to Poisson as indicated by $c^2_{a,1}$=0.96. If we replace the external source distribution by a hyper-exponential distribution with $c^2 = 4.0$ we obtain the results shown in Table 15.3. As expected, larger errors can be observed this time for the SCV of the arrival traffic. Interestingly, node 2 does not seem to be affected. This is because node 2 is fed by node 1 which is overloaded and hence reduces short-range correlations in the traffic stream.

Figure 15.5 shows the incoming traffic to node 1 as a function of the number of iterations in the fixed-point procedure for both kind of external sources. As can be observed, the fixed-point is reached after a very small number of iterations. This behavior has been typical for all queueing networks we have analyzed so far.

Kühn's nine-node network As a larger queueing network we evaluated a modified version of Kühn's nine-node network [27], as shown in Figure 15.6 (the numbers at the edges specify the routing probabilities). A similar network has been ex-

		analysis	simulation	rel. error
Output traffic	$\lambda_{netd,3}$	0.99	0.99	0.0%
	$c^2_{netd,3}$	0.45	0.69	-34.8%
Arrival traffic	$\lambda_{a,1}$	1.63	1.63	0.0%
	$c^2_{a,1}$	3.33	2.35	41.7%
	$\lambda_{a,2}$	1.33	1.33	0.0%
	$c^2_{a,2}$	0.79	0.79	0.0%
	$\lambda_{a,3}$	1.32	1.33	-0.8%
	$c^2_{a,3}$	0.35	0.65	-46.2%
Queue length	$\mathrm{E}[N_1]$	5.57	5.59	-0.4%
	$\mathrm{E}[N_2]$	3.30	3.16	4.4%
	$\mathrm{E}[N_3]$	2.38	2.76	-13.8%

Table 15.3: Results for the 3-node network with hyper-exponential source

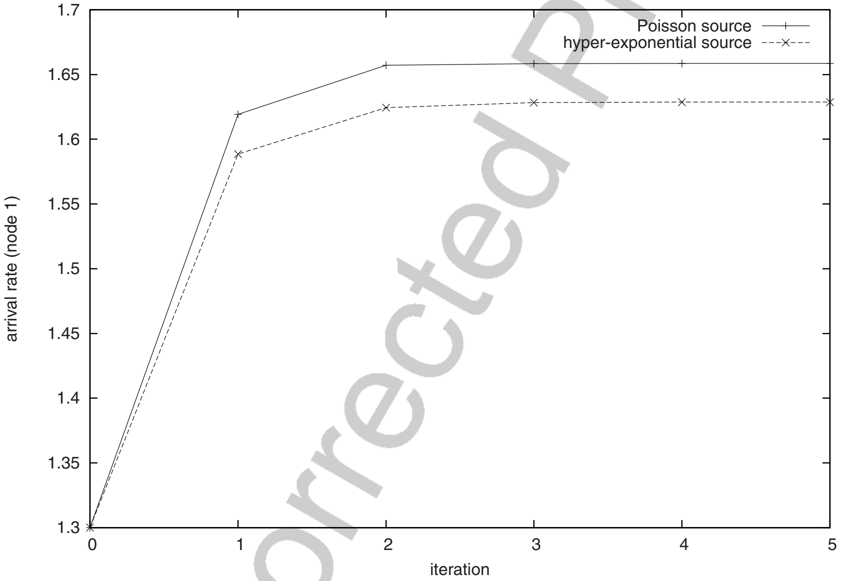


Fig. 15.5: Incoming traffic (arrival rate) at node 1 as a function of the number of iterations in the fixed-point procedure for the 3-node network

amined in [15, 48]. The external arrival rate to nodes 1–3 equals 0.8 and $c^2_{ext} = 4.0$. The service rate at each node is 1.0 (except for node 5 where $\mu_5 = 0.5$), and the SCV of all service processes is $c^2_s = 0.5$. All nodes have a finite queueing capacity of 25. Hence, without decomposition the underlying CTMC would comprise $2^3 \cdot (1+25\cdot 2)^9 \approx 1.86\times 10^{16}$ states. For all nodes, we observe excellent agreement with the simulation results.

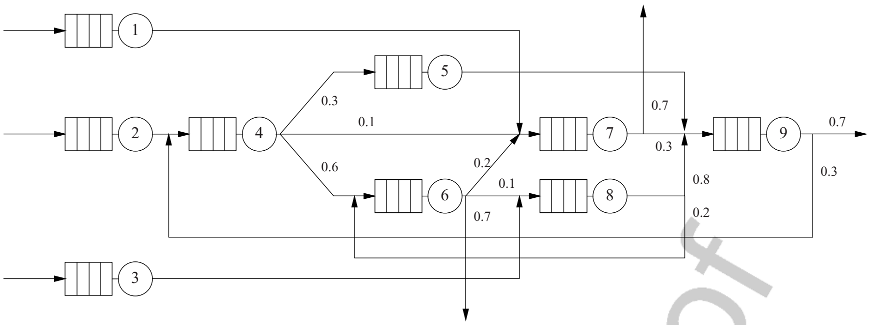


Fig. 15.6: Kühn’s nine-node network

Table 15.4 shows the results obtained by FiFiQueues and by simulation for the mean queue length and the offered load at each station. Note that the results for the (identical) nodes 1–3 are only stated once.

node		analysis	simulation	rel. error
1–3	E$[N_i]$	6.39	6.39	0.0%
	offered load	0.8	0.8	0.0%
4	E$[N_4]$	16.84	16.74	0.6%
	offered load	1.09	1.09	0.0%
5	E$[N_5]$	1.14	1.13	0.9%
	offered load	0.59	0.59	0.0%
6	E$[N_6]$	2.31	2.28	1.3%
	offered load	0.77	0.76	1.3%
7	E$[N_7]$	14.67	14.86	-1.3%
	offered load	1.04	1.04	0.0%
8	E$[N_8]$	6.36	6.63	-4.0%
	offered load	0.87	0.87	0.0%
9	E$[N_9]$	22.41	21.88	2.4%
	offered load	1.28	1.27	0.8%

Table 15.4: Results for the departure rates in Kühn’s nine-node network

15.5.2 Performance evaluation of a web server

In this section we will use FiFiQueues for the performance evaluation of a web server. The employed parameters in the models have been derived from measurements made at a test system.

This section is structured as follows. First, we describe the test system in Section 15.5.2.1. Then we present a QN model for a web server without disk access (cache only) in Section 15.5.2.2, followed of the model of a web server with disk access in Section 15.5.2.3. These two models are then combined to a model of a server group in Section 15.5.2.4. We compare the results obtained by analysis with the results obtained by simulation, and, where available, with the data collected at the test system.

15.5.2.1 Description of the test system

The test system consists of a computer running the Apache web server [2]. The server load is generated by two client systems that send HTTP/1.0 GET requests to the server in a 100 MBit Ethernet LAN. The request times as well as the sizes of the requested files have been extracted from traces (access logs) collected at the UC Berkeley Home IP Service [11] in 1996. For our tests we have used a part of the original trace file: it consists of 35541 requests for static files (i.e., pictures, HTML pages, etc.) sent over 4 hours by different users. This corresponds to a request rate of 2.468 requests per second. The SCV of the inter-request time is 1.2. The requested files have a mean size of 8510 bytes where the smallest file has a size of 2 bytes and the largest file a size of about 4.5 MBytes. The size distribution has a SCV as large as 26.8.

The web server of the test system has been configured to use not more than 150 server threads. This implies that the number of requests that can be processed concurrently is limited to 150. Since connection requests are not queued the clients will experience a connection rejection if they try to exceed this number. In addition, the request time-out has been set to 8 seconds. More details concerning the test system can be found in [25]; please note that the QN models presented in the following differ from the models discussed there.

15.5.2.2 Web server without disk access

For the first model, we assume that the server holds all requested files in the file cache and, as a consequence, no disk access is performed. This is a typical situation in intranets where the number of often requested files is limited. In this scenario the performance of the web server is only limited by the CPU, the main memory, and the network interface controller (NIC).

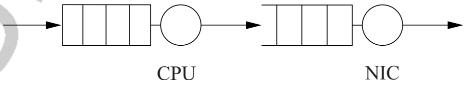


Fig. 15.7: QN model for the web server without disk access

We model the web server by two queueing stations in series as shown in Figure 15.7. Both stations have a finite queueing capacity; we comment on how the buffer capacity is chosen below. The first station is fed by an external source that represents the clients sending the HTTP requests. The SCV of 1.2 for the source is equal to the corresponding value of the trace file.

The first station models the CPU. Measurements at the test system have shown that the CPU of the test server is able to process up to around 1200 requests per second. We adopt this value for the service rate of the first queueing station. Concerning the SCV of the CPU's service time distribution, we observe that the CPU service time is dominated by the time to handle the HTTP protocol and by the management of the cache data structures. Since the NIC accesses the main memory via DMA (Direct Memory Access), the CPU service time exhibits nearly no dependency on the size of the requested file. Hence, we choose a (nearly) deterministic service time distribution with a SCV of 0.1. The second queue represents the NIC. Measurements have shown a network load between 90% and 95% for a response rate of 1100 responses per second. This leads to a NIC service rate of approximately 1200. For the SCV of the NIC's service time distribution, we assume a direct dependency of the service time on the file size and we set the SCV to 26.8, i.e., to the SCV of the file size distribution.

The most problematic aspect of the test system is the limitation to 150 simultaneously connected clients. This cannot be easily modeled by the FCFS-scheduling used by FiFiQueues. To approximate the limit, we have first analyzed the network at a request rate of 1500 requests per second. Using a Newton-iteration, we have determined the queueing capacity at which the total mean number of jobs in the network equals 150. The thus found buffer capacity of 106 has then been used for *all* other request rates (we have chosen the same capacity for both queues; the jobs are distributed evenly over both stations at high request rates).

Figure 15.8 shows the number of responses per second as a function of the number of requests sent per second as measured at the test system and as computed by FiFiQueues. Simulation results are not shown since they are nearly identical to the analytical results (relative error $< 1\%$). It shows that the QN model is able to predict the response rate quite well. The total mean response times are shown in Figure 15.9. The results are acceptable, but we can see that the model is not able to reproduce the sharp jump of the response time at 1000 requests/s. A model with more complex behavior, for example non-FCFS scheduling, would be required in order to obtain better results.

15.5.2.3 Web server with disk access

The second model assumes that all requested files have to be loaded from the disk of the server system. Measurements have shown that the test system only achieves a maximum response rate of 63.5 files/s at a CPU load of 9%. Clearly, the disk transfer is the bottleneck.

We model the influence of the disk access through an additional queueing station. Figure 15.10 shows the resulting model. The first station represents the CPU. For the SCV of the CPU service time, we have kept the value of 0.1 of the previous model.

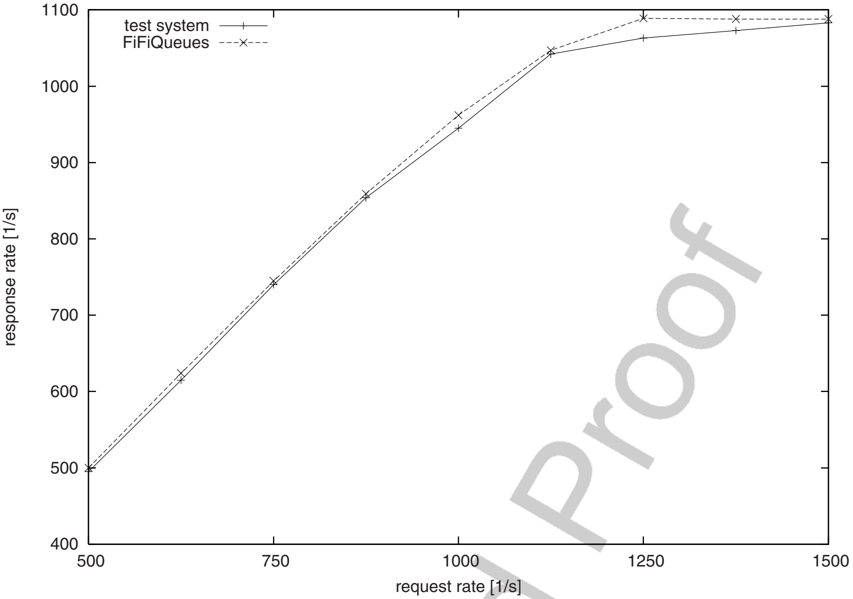


Fig. 15.8: Response rate as function of the request rate for the web server without disk access

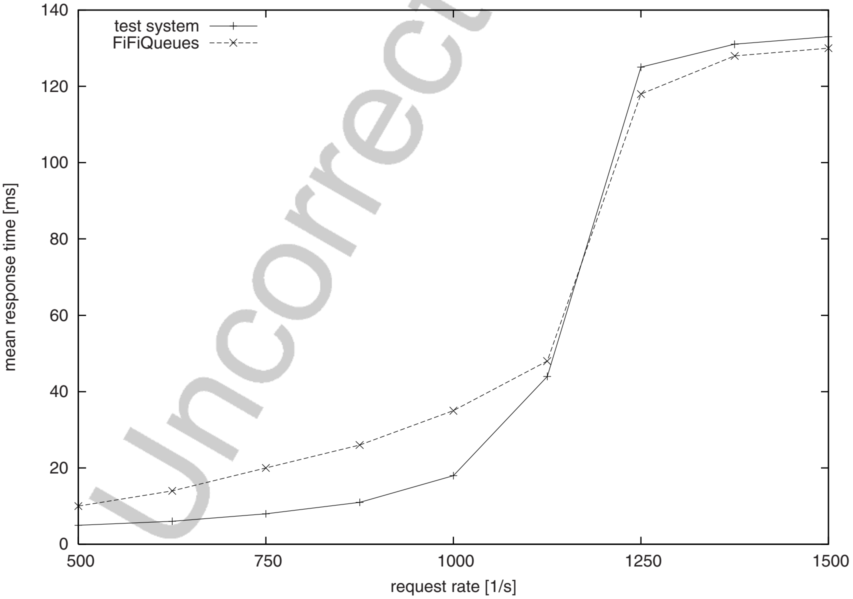


Fig. 15.9: Mean response time as function of the request rate for the web server without disk access

However, the service rate has now been set to a value of 706 ($= \frac{63.5}{0.09}$) to reflect the higher CPU demand of the single disk-based request. The service rate of the disk station has been set to 63.5. For the SCV, we have assumed a direct dependency of the service time on the size of the requested file *measured in blocks of 4 KBytes* since this corresponds to the organization of the data on the disk. This leads to a SCV of 16.5 instead of 26.8. The NIC in this model has the same service rate and SCV as in the previous model.

Again, the problem of the bounded number of simultaneously connected clients remains. Since the disk station clearly is the bottleneck, we have limited its queueing capacity to 150 while the CPU and the NIC station now have infinite queueing capacity. Note that, in spite of the large differences between the service rates, the CPU and the NIC should not be removed from the model since they have a small but measurable influence on the SCV of the traffic stream.

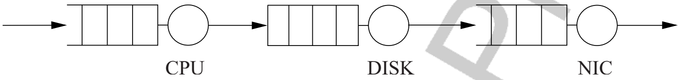


Fig. 15.10: QN model for the web server with disk access

Figure 15.11 shows the number of responses per second as a function of the number of requests sent per second as measured at the test system and as computed by FiFiQueues. Again, simulation results are not shown since their are nearly identical to the analytical results (relative error $< 1\%$). Again, it shows that the QN model is able to predict the response rate quite well. The total mean response times are shown in Figure 15.12. We observe that the QN model underestimates the response time, especially at request rates near to the maximum response rate of the disk. Our experiments with more complex QN models have shown that an improvement of the results cannot be easily achieved by using the type of queueing stations offered by FiFiQueues. For example, a more appropriate model would have to consider that the seek time of the disk becomes a significant part of the disk's response time at high file reqest rates since the disk has to reposition its read/write heads more often. Detailed models like the one presented in [39] simulate axial and rotational head positions, seek, rotation and transfer times, and provide separate submodels for the disk mechanism, the cache and the DMA engine.

15.5.2.4 Group of servers

In this section we evaluate a group of servers as shown in Figure 15.13. In our model, the client requests HTML pages from the main server of a web site. An HTML file refers to, on average, three other objects (company logo, images,. . .) that are also located on the main server. In addition, the HTML file refers to an object located on one of the five data servers. We assume that the HTML file and the three referred files located on the main server are frequently requested and, hence, the main

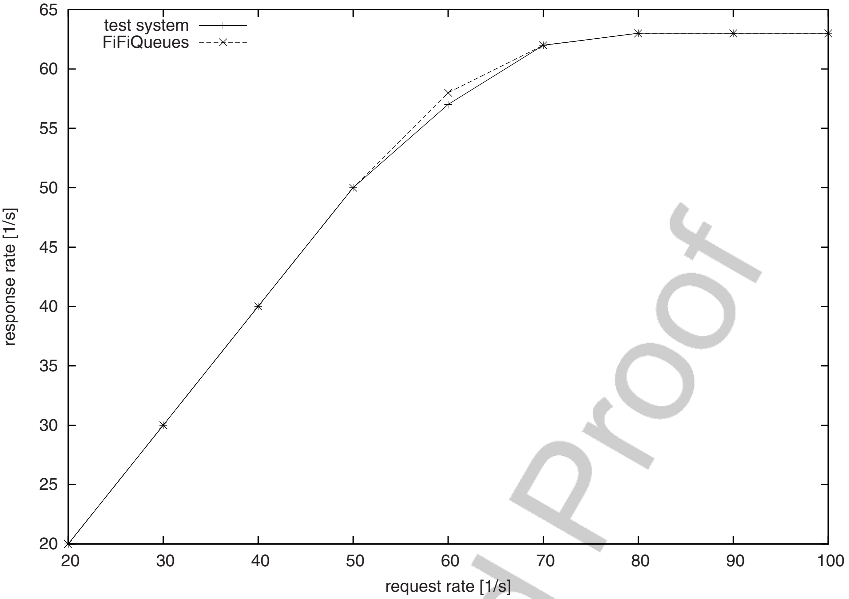


Fig. 15.11: Response rate as function of the request rate for the web server with disk access

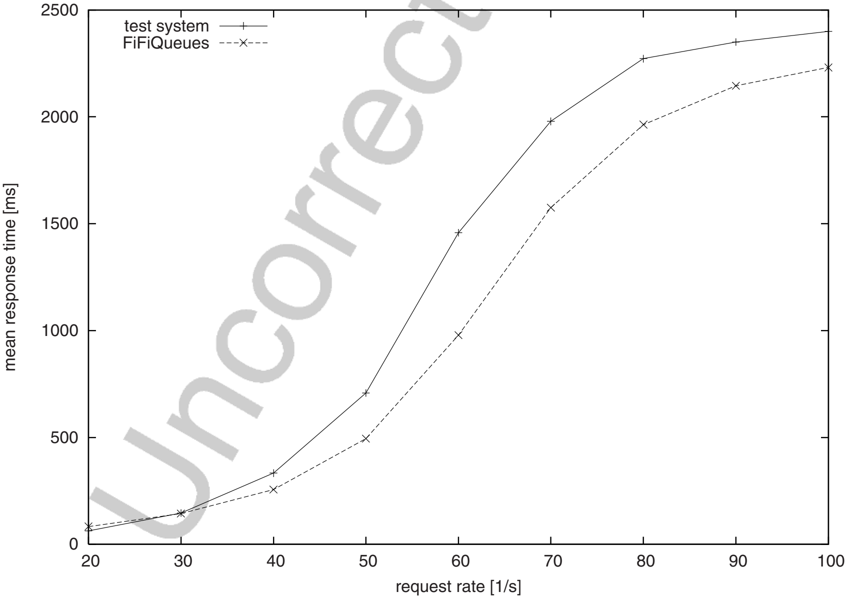


Fig. 15.12: Mean response time as function of the request rate for the web server with disk access

server mainly operates on the cache. Concerning the data servers, we assume that they store large amounts of infrequently requested files, for example files specific to the requesting user, media files, et cetera. The client uses the HTTP/1.0 protocol [35], i.e., the five files that constitute the requested HTML page are sequentially requested.

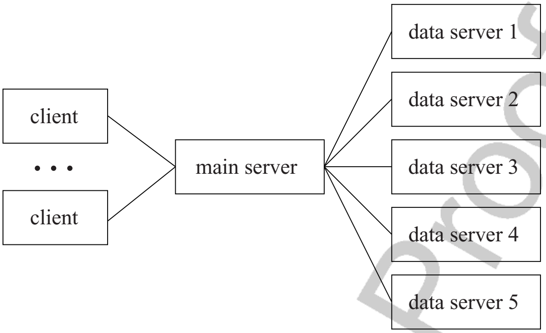


Fig. 15.13: Group of Web servers

The QN model is shown in Figure 15.14. The QN of the server without disk access (representing the main server) is combined with five copies of the QN of the server with disk access (representing the data servers). Jobs leaving the main server are fed back to it with a probability of 0.75, thus resulting in four visits to the main server in average. The jobs finally leaving the main server are distributed evenly on the data servers. The service processes and the capacities of the stations remain unchanged.

We have evaluated the QN model by FiFiQueues and by simulation. The results for the response rate (for one data server) and the mean response time are shown in Figure 15.15 and, respectively, Figure 15.16. The vertical bars in the latter show the 95%-confidence intervals of the simulation results. FiFiQueues provides good results for request rates smaller than 250. At larger request rates, FiFiQueues overestimates the losses in the main server because it ignores the correlations caused by the feedback. As a consequence, the load of the data servers is underestimated which leads to a smaller mean response time in comparison with the results obtained by simulation. To make this very clear: the differences we observe here show shortcomings of our analysis approach, as for both curves, the same model is being used.

Table 15.5 shows the runtimes (in seconds) of the FiFiQueues algorithm and of the discrete-event simulation for the evaluation of the server group model with various request rates. For FiFiQueues, we have recorded the runtimes for two different implementations of the finite queue analysis. The original implementation uses a Gauss-Seidel iteration, whereas the latest version uses the Cyclic Reduction method [3]. As observed in [40], the runtime of the Gauss-Seidel iteration increases

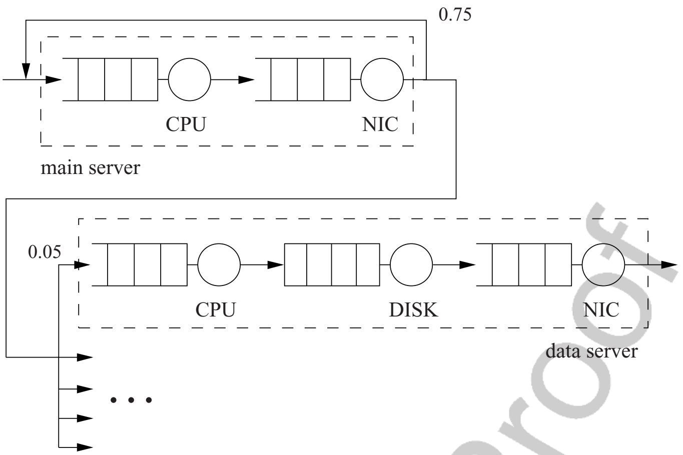


Fig. 15.14: QN model for the server group

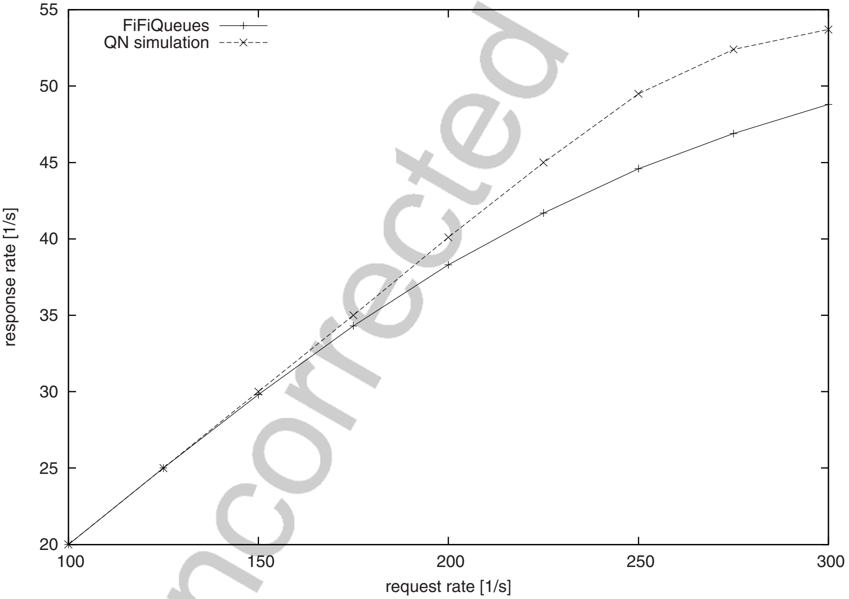


Fig. 15.15: Response rate as function of the request rate for the server group

	FiFiQueues		
request rate	Gauss-Seidel	Cyclic Red.	simulation
100	7	2	11
200	15	3	11
300	19	3	11

Table 15.5: Runtimes (in seconds) for the evaluation of the server group model

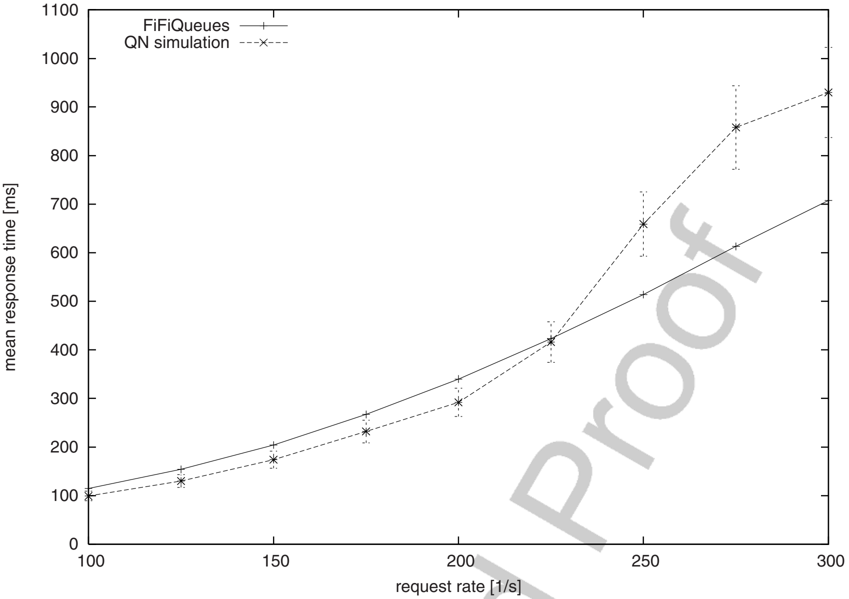


Fig. 15.16: Mean response time as function of the request rate for the server group

with the load of the stations. The Cyclic Reduction method is clearly faster than the Gauss-Seidel iteration and the simulation.

15.5.3 Summary

In this section, we have evaluated the performance of the FiFiQueues algorithm. Our experiments have shown that FiFiQueues provides very good results for important performance measures, like mean queue length, if the involved arrival times in the queueing network are hypo-exponentially or nearly (negative-)exponentially distributed. In such situations, we can generally expect relative errors less than 5%, even if the network has a complex structure. In case of hyper-exponential arrival processes, especially in queueing networks with feedback, relative errors up to 10%, rarely up to 20%, have been observed.

15.6 Summary and conclusions

In this chapter we have presented an overview of decomposition-based analysis techniques for large open queueing networks. We first presented the decomposition-based approach in general terms, without referring to any particular model class, and proposed a general fixed-point iterative solution method for it. We concretized this framework by describing the well-known QNA method, as proposed by Whitt in the early 1980s, in that context, before describing our FiFiQueues approach. It should be noted that the work on FiFiQueues has been performed by a group of people over the last (almost) 15 years. To keep this chapter self-contained, we have added appendices on various underlying building blocks. In addition to an extensive evaluation with generally very favorable results for FiFiQueues, we also present a theorem on the existence of a fixed-point solution for FiFiQueues (which has not been published before).

In [40], we have also experimented with three-moment traffic descriptors, as well as with traffic descriptors taking into account correlations in the traffic streams. However, in our experiments, the three-moment descriptors have not significantly improved the results for queueing networks with feedback in comparison to the two-moment descriptors used by FiFiQueues. Since three-moment descriptors considerably increase the runtime of the analysis, we currently refrain from using them. Incorporating correlations in the traffic descriptors does hold promise, however, this should be investigated further before it can be made into a daily practice.

15.7 Appendix: Jackson queueing networks

The simplest open queueing networks allowing feedback are the so-called Jackson queueing networks (JQNs). Their analytical performance evaluation was developed by J.R. Jackson [20] in the 1950s.

15.7.1 Model class

In JQNs, all nodes are assumed to be infinite-buffer M|M|1 queues with the First-Come-First-Served (FCFS) service discipline. In many modeling applications, the restriction to Poisson arrival and service processes cannot be justified.

15.7.2 Traffic descriptor

In JQNs, all traffic processes (including the external arrival processes) are assumed to be Poisson, hence a sufficient traffic descriptor only contains the arrival rate λ of the traffic stream, denoted as $\langle\lambda\rangle$.

15.7.3 Superposition of traffic streams

Merging two (possibly dependent) traffic streams does not necessarily yield a new Poisson stream. However, it can be shown that the nodes of a JQN still can be described by M|M|1 queues even when traffic merging occurs. Thus, to merge n traffic streams specified by $\langle\lambda_1\rangle, \ldots, \langle\lambda_n\rangle$ into one traffic stream $\langle\lambda\rangle$, one simply adds the rates:

$$\lambda = \sum_{i=1}^{n} \lambda_i.$$

15.7.4 Splitting traffic streams

The Markovian splitting of a Poisson stream $\langle\lambda\rangle$ again results in n Poisson streams. Let $p_1, \ldots, p_n$ be the splitting probabilities, then the resulting streams $\langle\lambda_1\rangle, \ldots, \langle\lambda_n\rangle$ are given by

$$\lambda_i = p_i \cdot \lambda, \qquad i = 1, \ldots, n.$$

15.7.5 Servicing jobs

Let $\langle\lambda_A\rangle$ be the arrival traffic descriptor of the node, and μ its service rate. We require that $\lambda_A < \mu$, otherwise the station is not stable. Burke [7] proved that the departure process for a stable single server M|M|1 queue is a Poisson process with rate λ_A, hence, the departure process can be described as $\langle\lambda_D\rangle$ with $\lambda_D = \lambda_A$.

15.7.6 Node performance

Let $\langle\lambda_A\rangle$ be the arrival traffic descriptor, and μ the service rate of the node. Then, $\rho = \lambda_A/\mu$ is the utilization of the node. Since the node is an M|M|1 queue, the steady-state probability p_j to find j customers in the queue can be easily derived from the underlying birth-death Markov chain [16]:

$$p_j = (1-\rho)\rho^j, \qquad j = 0, 1, \ldots$$

Having computed the steady-state probabilities, quantities like the expected number of jobs in the queueing station E[N] can be calculated as

$$\mathrm{E}[N] = \sum_{j=0}^{\infty} j \cdot p_j = \frac{\rho}{1-\rho}.$$

Then, Little's law can be applied to compute the expected waiting time $\mathrm{E}[W]$.

Similarly, higher moments of measures can be computed too, e.g., the variance of the number of customers in the node:

$$\mathrm{Var}[N] = \sum_{j=0}^{\infty} (j - \mathrm{E}[N])^2 \cdot p_j = \frac{\rho}{(1-\rho)^2}.$$

15.7.7 Network-wide performance

Since no losses occur and all nodes are required to be stable, the total throughput λ_{thr} of the network, i.e., the average number of customers passing through the network per time unit, is simply the sum of arrival rates $\lambda_{ext,i}$ of the external arrival processes:

$$\lambda_{thr} = \sum_{i=1}^{n} \lambda_{ext,i}$$

where $\langle \lambda_{ext,i} \rangle$ is the external traffic arriving at node i and n is the number of nodes. Other performance measures may be derived from the node performance measures. If $\lambda_{A,i}$ is the total amount of traffic arriving at node i, the expected number of visits $\mathrm{E}[V_i]$ of a customer at node i is given by [49, Eq. (77)]:

$$\mathrm{E}[V_i] = \lambda_{A,i} / \lambda_{thr}.$$

The expected total sojourn time $\mathrm{E}[T_{total}]$, i.e., the time a customer spends in the network, defined as the sum of the expected sojourn times $\mathrm{E}[T_i]$ at each node i, thus equals

$$\mathrm{E}[T_{total}] = \sum_{i=1}^{n} \mathrm{E}[T_i] = \sum_{i=1}^{n} \mathrm{E}[V_i] \left(\frac{1}{\mu_i} + \mathrm{E}[W_i] \right).$$

Since the total number of customers N_{total} in the network is the sum of customers present in each queueing station, we have

$$\mathrm{E}[N_{total}] = \sum_{i=1}^{n} \mathrm{E}[N_i],$$

where $\mathrm{E}[N_i]$ is the expected number of jobs in node i.

15.7.8 Complexity

If $\Gamma = (r_{ij})$ is the routing matrix, the traffic $\langle \lambda_{A,i} \rangle$ arriving at node i is given by the so-called*first-order traffic equation*:

$$\lambda_{A,i} = \lambda_{ext,i} + \sum_{j=1}^{n} \lambda_{D,j} \cdot r_{ji}.$$

Since $\lambda_{D,i} = \lambda_{A,i}$, the traffic equations form a system of linear equations which can be expressed in vector/matrix notation as $\lambda_{\mathrm{A}} = \lambda_{\mathrm{ext}} + \lambda_{\mathrm{A}} \cdot \Gamma$, or, after transformation, as

$$\lambda_{\mathrm{A}} = \lambda_{\mathrm{ext}}(\mathrm{I} - \Gamma)^{-1}.$$

Thus, to find λ_{A} we solve the linear system

$$\lambda_{\mathrm{A}}(\mathrm{I} - \Gamma) = \lambda_{\mathrm{ext}}.$$

This system of equations can be solved by direct methods like Gaussian elimination, resulting in a time complexity of $O(n^3)$, or by iterative methods like Gauss-Seidel. This implies that, due to the linearity of the involved equations, the analysis of JQNs does not require the fixed-point iteration described in Section 15.2 (although, if an iterative solver is used to solve the linear system, the fixed-point iteration can be regarded as hidden in the solver).

For very large networks, we can make use of the fact that the routing matrix typically is a sparse matrix. In this way, the time complexity of an iterative solver such as Gauss-Seidel can be reduced to about $O(c \cdot n)$ where c is the average number of outgoing connections per station.

The expressions given in Section 15.7.6 for the node performance measures can be computed in constant time for each node. For the network performance, most results require summation over the number of nodes in the network which yields a time complexity of $O(n)$.

15.8 Appendix: MAPs, PH-distributions and QBDs

In this appendix we introduce the fundamental mathematical structures and notation used throughout this chapter. We begin with an important class of stochastic processes, the Markovian Arrival Processes (MAP) in Section 15.8.1. Phase-type (PH) renewal processes, which can be seen as special cases of MAPs, are introduced in Section 15.8.2. The queueing processes that we have discussed in Section 15.4 have underlying Markov chains that belong to the well-known class of continuous-time Quasi-Birth-and-Death (QBD) processes. We give the formal definition of QBD processes as well as methods to compute their steady-state solution. We first discuss infinite QBDs in Section 15.8.3 and continue with finite QBDs in Section 15.8.4.

15.8.1 Markovian Arrival Processes (MAPs)

15.8.1.1 Definition and notation

Markovian Arrival Processes (MAPs) [29, 30, 36] belong to the general class of point processes and can be seen as special cases of Matrix Exponential Point Processes (which, in turn, form a subset of the class of Semi-Markov Processes [16]). MAPs cover many interesting processes including the Markov-Modulated Poisson Processes (MMPPs) [10] and the phase-type (PH) renewal processes (see below).

A MAP can be described by a finite irreducible continuous-time Markov chain (CTMC) with generator matrix Q where some transitions are "marked". Every time when the process passes through such a marked transition an event is triggered. The time instants of these events form the point process. We follow the notation of [31] and split the generator matrix into two matrices Q_0 and Q_1 as follows:

$$\mathrm{Q}_0 = \begin{pmatrix} -q_1 & q_{12} & \cdots & q_{1m} \\ q_{21} & -q_2 & \cdots & q_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ q_{m1} & q_{m2} & \cdots & -q_m \end{pmatrix}, \quad \mathrm{Q}_1 = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{pmatrix},$$

with $\mathrm{Q}_0 + \mathrm{Q}_1 = \mathrm{Q}$ where $q_i = a_{ii} + \sum_{j=1, j\neq i}^{m} (q_{ij} + a_{ij})$. The elements of the matrix Q_1 give the transition rates of the marked transitions.[1] In the following, we denote a MAP by the pair $(\mathrm{Q}_0, \mathrm{Q}_1)$ and call m the *size* of the MAP.

15.8.1.2 Characteristics

Some general results of the Markov-modulated Poisson process [10] can be easily adapted to the MAP. In order to compute the behavior of a MAP $(\mathrm{Q}_0, \mathrm{Q}_1)$ we first need to choose the initial probability vector p of the MAP. In analogy to phase-type renewal processes we start the MAP at an "arbitrary" arrival epoch by choosing

$$\mathrm{p} = \frac{1}{\pi \mathrm{Q}_1 1} \pi \mathrm{Q}_1,$$

where π is the steady-state probability vector of the MAP, i.e., $\pi(\mathrm{Q}_0 + \mathrm{Q}_1) = 0$. The thus-obtained process is said to be *interval-stationary*. The inter-arrival time distribution function of the interval-stationary process is given by

$$F(t) = 1 - \mathrm{p} \exp(\mathrm{Q}_0 t) 1, \tag{15.1}$$

[1] This definition allows the following interpretation of the matrices Q_0 and Q_1: passing through a transition given as entry of Q_1 triggers the generation of *one* event. Batch Markovian Arrival Processes (BMAPs) generalize this viewpoint by introducing matrices Q_i with $i > 1$ whose entries describe transitions with batch arrivals of size i.

which leads to the following expression for the kth moment of the inter-arrival time:

$$\mathrm{E}[T^k] = k!\mathrm{p}(-\mathrm{Q}_0)^{-(k+1)}\mathrm{Q}_1 1. \tag{15.2}$$

Hence, the first moment of the inter-arrival time is given by

$$\mathrm{E}[T] = \frac{1}{\pi\mathrm{Q}_1 1}\pi\mathrm{Q}_1(-\mathrm{Q}_0)^{-2}\mathrm{Q}_1 1.$$

This equation can be further simplified by using the equations $\pi\mathrm{Q}_1 = -\pi\mathrm{Q}_0$ and $\mathrm{Q}_1 1 = -\mathrm{Q}_0 1$ which follow from the definition of π, respectively, from the fact that $\mathrm{Q}_0 + \mathrm{Q}_1$ is a stochastic matrix. We find that the arrival rate λ of a MAP (the inverse of the first moment) is

$$\lambda = \pi\mathrm{Q}_1 1$$

which yields

$$\mathrm{E}[T^k] = \frac{k!}{\lambda}\pi(-\mathrm{Q}_0)^{-(k-1)}1.$$

Let T_i be the time between the ith and the $(i+1)$st arrival in a MAP. Then, the autocovariance function $R(k)$ for T_1 and T_{k+1} with $k \geq 1$ is given by

$$\begin{aligned} R(k) &= \mathrm{E}\left[(T_1 - \mathrm{E}[T_1])(T_{k+1} - \mathrm{E}[T_{k+1}])\right] \\ &= \mathrm{p}(-\mathrm{Q}_0)^{-2}\mathrm{Q}_1\left\{\left[(-\mathrm{Q}_0)^{-1}\mathrm{Q}_1\right]^{k-1} - 1\mathrm{p}\right\}(-\mathrm{Q}_0)^{-1}1. \end{aligned}$$

The limiting index of dispersion I of a MAP is given by [18]

$$I = \lim_{t\to\infty}\frac{\mathrm{Var}[N(t)]}{E[N(t)]} = 1 + 2\left(\lambda - \frac{1}{\lambda}\pi\mathrm{Q}_1(\mathrm{Q}_0 + \mathrm{Q}_1 + 1\pi)^{-1}\mathrm{Q}_1 1\right),$$

where $N(t)$ is the counting process of the MAP.

15.8.1.3 Superposition and Markovian splitting

The class of MAPs is closed under superposition and Markovian splitting. The superposition of two MAPs $(\mathrm{A}_0, \mathrm{A}_1)$ and $(\mathrm{B}_0, \mathrm{B}_1)$ is a new MAP $(\mathrm{C}_0, \mathrm{C}_1)$ with

$$\mathrm{C}_0 = \mathrm{A}_0 \oplus \mathrm{B}_0, \quad \mathrm{C}_1 = \mathrm{A}_1 \oplus \mathrm{B}_1,$$

where $\mathrm{L} \oplus \mathrm{M} = \mathrm{L} \otimes \mathrm{I} + \mathrm{I} \otimes \mathrm{M}$, and $\otimes$ is the Kronecker product operator (also known as tensor or matrix direct product operator).

The Markovian splitting of a MAP $(\mathrm{A}_0, \mathrm{A}_1)$ with probability r gives two MAPs $(\mathrm{B}_0, \mathrm{B}_1)$ and $(\mathrm{C}_0, \mathrm{C}_1)$ with

$$\begin{aligned} (\mathrm{B}_0, \mathrm{B}_1) &= (\mathrm{A}_0 + (1-r)\mathrm{A}_1, r\mathrm{A}_1), \\ (\mathrm{C}_0, \mathrm{C}_1) &= (\mathrm{A}_0 + r\mathrm{A}_1, (1-r)\mathrm{A}_1). \end{aligned}$$

15.8.1.4 Markov-Modulated Poisson Processes (MMPPs)

The MMPP is the doubly stochastic Poisson process whose arrival rate depends on the state of an irreducible Markov process. Thus, MMPPs can be seen as MAPs where the matrix Q_1 is restricted to the form

$$\begin{pmatrix} a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_{mm} \end{pmatrix}.$$

15.8.2 Phase-type (PH) renewal processes

15.8.2.1 Definition and notation

A continuous phase-type renewal process can be seen as a special MAP $(\mathrm{A}, \mathrm{A}^0\alpha)$ where A^0 is a $n \times 1$ column vector with entries and α is a $1 \times n$ row probability vector. Consequently, it holds $\mathrm{A}^0 = -\mathrm{A}1$.

We adopt the notation of [37] and denote PH renewal processes by the pair (α, A) which can be interpreted as follows: the $n \times n$ matrix A describes the transitions from the n transient states of a CTMC with $n+1$ states. The last state $n+1$ is an absorbing state and any transition (given by A^0) from the transient state to the absorbing state will trigger an arrival. After the arrival, the process will restart in the transient state i with probability α_i. Furthermore, PH inter-event time distributions form a dense subset of all distributions with support on $[0;\infty)$, i.e., any distribution can be approximated arbitrarily closely by a PH distribution [23].

15.8.2.2 Inter-event time characteristics

Obviously, the vector α of the PH renewal process (α, A) is identical to the interval-stationary probability vector p of the corresponding MAP. Hence, expressions for the distribution function of the inter-event time and the k-th moment directly follow from Equations (15.1) and (15.2) and we have

$$F(t) = 1 - \alpha \exp(\mathrm{A}t)1,$$

respectively

$$\mathrm{E}[T^k] = k!\alpha(-\mathrm{A})^{-k}1.$$

Note that the matrix A is nonsingular, so that all moments are finite. From this follows that the MAP $(\mathrm{Q}_0, \mathrm{Q}_1)$ and the PH renewal process $(\mathrm{p}, \mathrm{Q}_0)$ have the same inter-event time distribution.

15.8.2.3 Superposition and Markovian splitting

The superposition of two PH renewal processes (α,A) and (β,B) is a MAP $(\mathrm{C}_0,\mathrm{C}_1)$ with

$$\mathrm{C}_0 = \mathrm{A}\oplus\mathrm{B}, \quad \mathrm{C}_1 = \mathrm{A}^0\alpha\oplus\mathrm{B}^0\beta.$$

Note that the class of PH renewal processes is not closed under superposition.

The Markovian splitting of a PH renewal processes (α,A) with probability r gives two PH renewal processes $(\alpha,\mathrm{A}+(1-r)\mathrm{A}^0\alpha)$ and $(\alpha,\mathrm{A}+r\mathrm{A}^0\alpha)$.

15.8.3 Infinite QBDs

This section is based on Chapter 4 of [38]. Note that we use a simplified notation.

15.8.3.1 Definition

QBD processes [37] can be described as a generalization of the queueing process of M|M|1 queueing stations. In the underlying Markov chain of such a queue we can identify an infinite number of states where state i describes that i jobs are in the system. The transition from state i to $i+1$ resp. from $i+1$ to i is marked by the arrival rate resp. the service rate of the queueing station.

In QBDs, these states are replaced by so-called *levels*: level i still stands for i jobs in system but in QBDs each level may consist of more than one state. Usually, a two-dimensional addressing scheme is used for the states where (i,j) addresses state j of level i. Note that in the QBD the number of levels is unbounded whereas the number of states per level is required to be finite. Moreover, the levels $1,2,\ldots$ (the *repeating levels*) have to contain the same number of states N. Level 0 is called *boundary level* and may contain a different number of states N_0.

In QBDs, two adjacent levels i and $i+1$ are not connected by one single transition. Instead, arbitrary transitions between the states of two adjacent levels and between states of the same level are allowed. Consequently, the transition rates are specified by matrices:

- the entry (i,j) of the $N_0\times N$ matrix $\mathrm{B}_{0,1}$ gives the transition rate from state $(0,i)$ to state $(1,j)$. The opposite direction (from level 1 to level 0) is given similarly by the $N\times N_0$ matrix $\mathrm{B}_{1,0}$.
- the entry (i,j) of matrix A_0 gives the transition rate from state (l,i) to state $(l+1,j)$ where $l=1,2,\ldots$. The opposite direction (from level $l+1$ to l) is given by matrix A_2. Both matrices are of size $N\times N$.
- transitions inside level 0 are specified by the $N_0\times N_0$ matrix $\mathrm{B}_{0,0}$. Entry (i,j) gives the transition rate from state $(0,i)$ to state $(0,j)$. Correspondingly, transitions inside repeating level l (with $l=1,2,\ldots$) are specified by the $N\times N$ matrix A_1.

As can be seen, all repeating levels have a similar transition structure. The above described matrices directly lead to the generator matrix of the QBD Markov chain. If we sort the states lexicographically, i.e., in the sequence

$$(0,1),\ldots,(0,N_0),(1,1),\ldots,(1,N),(2,1),\ldots$$

we obtain the tri-diagonal block generator matrix Q of infinite size:

$$\mathrm{Q} = \begin{pmatrix} \mathrm{B}_{0,0} & \mathrm{B}_{0,1} & & & \\ \mathrm{B}_{1,0} & \mathrm{A}_1 & \mathrm{A}_0 & & \\ & \mathrm{A}_2 & \mathrm{A}_1 & \mathrm{A}_0 & \\ & & \mathrm{A}_2 & \mathrm{A}_1 & \mathrm{A}_0 \\ & & & \ddots & \ddots & \ddots \end{pmatrix}. \tag{15.3}$$

15.8.3.2 Steady-state solution

The infinite steady-state probability vector v of the QBD Markov chain with generator matrix Q fulfills the global balance equation

$$\mathrm{v}\cdot\mathrm{Q} = 0, \tag{15.4}$$

and the normalization condition

$$\mathrm{v}\cdot\mathbf{1} = 1. \tag{15.5}$$

In the following we write v_0 for the vector $(v_1,\ldots,v_{N_0})$ which contains the steady-state probabilities for the states of level 0 and we write v_i for the vector $(v_{N_0+1+(i-1)\cdot N},\ldots,v_{N_0+i\cdot N})$ which contains the steady-state probabilities of level $i = 1,2,\ldots$. With this notation we can rewrite Equations (15.4) and (15.5) as

$$\mathrm{v}_0\mathrm{B}_{0,0} + \mathrm{v}_1\mathrm{B}_{1,0} = 0, \tag{15.6}$$

$$\mathrm{v}_0\mathrm{B}_{0,1} + \mathrm{v}_1\mathrm{B}_{1,1} + \mathrm{v}_2\mathrm{A}_2 = 0, \tag{15.7}$$

$$\mathrm{v}_\mathrm{i}\mathrm{A}_0 + \mathrm{v}_{\mathrm{i}+1}\mathrm{A}_1 + \mathrm{v}_{\mathrm{i}+2}\mathrm{A}_2 = 0, \qquad \text{for } i = 1,2,\ldots, \tag{15.8}$$

$$\sum_{i=0}^{\infty}\mathrm{v}_\mathrm{i}\mathbf{1} = 1. \tag{15.9}$$

The regular structure of Equation (15.8) is the key to the efficient solution of the QBD. Two different classes of solution techniques can be distinguished: matrix-geometric solution methods and transform methods. We will describe them briefly in the following.

15.8.3.3 Matrix-geometric solution methods

The main idea in this class of solution methods is that the solution vector v has a matrix-geometric form, i.e., there exists a matrix R of size $N \times N$ with

$$\mathrm{v_i} = \mathrm{v_1 R}^{i-1}, \qquad i = 1,2,\ldots \tag{15.10}$$

In [37], it is shown that R is the entry-wise smallest non-negative solution of the quadratic matrix equation

$$\mathrm{A_0} + \mathrm{RA_1} + \mathrm{R^2A_2} = 0. \tag{15.11}$$

The methods belonging to this class of solution methods try to solve this equation as efficiently as possible. Once R has been determined, the complete stationary vector v can be computed using Equations (15.6)–(15.10). Examples for such methods are the Successive Substitution method [37], the Logarithmic Reduction method [28] and its improvement [34].

15.8.3.4 Transform methods

Unlike the matrix-geometric solution methods the transform methods do not aim to directly solve Equation (15.11). Instead, they first transform the problem into some other domain in order to derive the solution of Equation (15.8). Three well known methods belonging to this class are the Cyclic Reduction method [4], the Invariant Subspace method [1], and the Spectral Expansion method [37, 8].

15.8.4 Finite QBDs

15.8.4.1 Definition

Similar to infinite QBDs, finite QBDs can be seen as the generalization of the queueing process of a bounded M|M|1|K queue. Finite QBD processes result in QBD Markov chains with a finite number $K+1$ of levels, hence two boundary levels can be identified: the *lower boundary level* 0 and the *upper boundary level* K.

In the following we will only treat a quite restricted class of finite QBDs that is sufficient for the queueing process discussed in this chapter: The upper boundary level has the same number of states N as the repeating levels 1 through $K-1$. Additionally, the transition rates between levels $K-1$ and K are the same as between the repeating levels — only one new matrix C is introduced that specifies the transition rates inside level K. The finite generator matrix of the QBD Markov chain then has the following form:

$$Q = \begin{pmatrix} B_{0,0} & B_{0,1} & & & & & \\ B_{1,0} & A_1 & A_0 & & & & \\ & A_2 & A_1 & A_0 & & & \\ & & A_2 & A_1 & A_0 & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & A_2 & A_1 & A_0 \\ & & & & & A_2 & C \end{pmatrix}. \tag{15.12}$$

15.8.4.2 Steady-state solution

We search the finite steady-state probability vector v of the QBD Markov chain with generator matrix Q that fulfills the global balance equation

$$\mathrm{v} \cdot \mathrm{Q} = 0 \tag{15.13}$$

and the normalization condition

$$\mathrm{v} \cdot \mathbb{1} = 1. \tag{15.14}$$

As in the infinite case, we partition the vector v into subvectors v_i where $\mathrm{v}_0 = (v_1, \ldots, v_{N_0})$ and $\mathrm{v_i} = (v_{N_0+1+(i-1)\cdot N}, \ldots, v_{N_0+i\cdot N})$, for $i = 1, \ldots, K$. It is important to note that the solution of Equation (15.13) is uncritical with respect to space complexity. Due to the special structure of the Markov chain it is not necessary to hold the whole matrix Q in memory but only the matrices $B_{0,0}$, $B_{1,0}$, $B_{0,1}$, $B_{1,1}$, A_0, A_1, A_2 and C. In terms of these matrices, Equation (15.13) becomes:

$$\mathrm{v_0 B_{0,0}} + \mathrm{v_1 B_{1,0}} = 0 \tag{15.15}$$

$$\mathrm{v_0 B_{0,1}} + \mathrm{v_1 B_{1,1}} + \mathrm{v_2 A_2} = 0 \tag{15.16}$$

$$\mathrm{v_i A_0} + \mathrm{v_{i+1} A_1} + \mathrm{v_{i+2} A_2} = 0, \qquad \text{for } i = 1, \ldots, K-2, \tag{15.17}$$

$$\mathrm{v_{K-1} A_0} + \mathrm{v_K C} = 0 \tag{15.18}$$

Since Q is of finite size, Equation (15.13) can be solved by an ordinary Gauss-Seidel-iteration which performs very efficiently due to the band-structure of Q. More sophisticated algorithms have been developed on the basis of the solution methods for infinite QBDs; most of the algorithms presented in Section 15.8.3 have been extended to the treatment of finite QBDs.

In addition to these methods some authors have developed solution methods especially adapted to QBDs arising from PH|PH|1|K queues (see Section 15.4). Such methods are the Bocharov-Naoumov method [6], two methods proposed by Chakravarthy and Neuts in [9], and the Cyclic-Reduction method [3].

15.9 Appendix: Existence of the fixed point

In general, it is not known for the fixed-point iteration algorithm described in Section 15.2 whether the searched fixed point exists, is unique or will be reached. However, some intermediate results are available for FiFiQueues which we will present here. In the following we give a proof that the fixed point exists for a modified version of the original FiFiQueues algorithm.

15.9.1 Notation and Brouwer's theorem

Given a queueing network with n stations, we define $D \subset \mathbb{R}^{2n}$ where the tuple

$$(\langle \lambda_{a,1}, c^2_{a,1} \rangle, \ldots, \langle \lambda_{a,n}, c^2_{a,n} \rangle) \in D$$

gives for each node $i \in \{1, \ldots, n\}$ the traffic descriptor $\left\langle \lambda_{a,i}, c^2_{a,i} \right\rangle$ of its arrival traffic. Then the operations performed by FiFiQueues during step $k+1$ of the fixed-point iteration can be expressed as a function $H : D \to D$ [47] which computes from the traffic descriptor d^k obtained from step k the new traffic descriptor d^{k+1}, that is,

$$\mathrm{d}^{k+1} = H(\mathrm{d}^k),$$

where d^0 is the initial traffic descriptor used in the iteration. We use the *Brouwer fixed-point theorem* [45] to prove the existence of the fixed point for the function H. It states:

Let $D \subset \mathbb{R}^m$ be a non-empty, closed, convex, and bounded set, and $H : D \to D$ continuous. Then H has a fixed point.

A first proof of the existence of the fixed point has been discussed in [47] for special service processes. The proof presented in the following applies to arbitrary PH renewal service processes. We first show in Section 15.9.2 that the requirements to the set D are met. The continuity of H is shown in Sections 15.9.3–15.9.5.

15.9.2 Properties of D

Lower and upper bounds for the arrival rate $\lambda_{a,i}$ of a node i exist. It holds that

$$0 \leq \lambda_{a,i} \leq \lambda_{max,i}$$

where the upper bound $\lambda_{max,i}$ is the maximum arrival rate that will only be reached if all queueing stations operate with a load of 100%. It is given by

$$\lambda_{\max} = \lambda_{\text{ext}}(\mathrm{I} - \Gamma)^{-1},$$

where Γ is the routing matrix and $\lambda_{ext,i}$ is the rate of the external traffic arriving at node i. As previously explained, FiFiQueues limits the squared coefficient of variation to $\frac{1}{10}$ to prevent the generation of PH distributions with more than 10 states. Originally, no upper bound is provided for the coefficients but we can safely define

$$c^2_{a,i} := \min(c^2_{max}, c^2_{a,i}), \qquad i = 1,\dots,n,$$

with $c^2_{max} = 1000$ without affecting the analysis. We thus obtain that D is the non-empty, closed and convex interval

$$\left[\left(\langle 0, \tfrac{1}{10}\rangle, \dots, \langle 0, \tfrac{1}{10}\rangle\right), \left(\langle \lambda_{max,1}, c^2_{max}\rangle, \dots, \langle \lambda_{max,1}, c^2_{max}\rangle\right)\right] \subset \mathbb{R}^{2n},$$

as required.

15.9.3 Continuity of H

The function H performs for each node the following operations to compute the traffic descriptors for the next iteration in the algorithm:

1. the service operation;
2. the traffic splitting;
3. the traffic merging.

The traffic merging step is a function of the traffic descriptors generated during the splitting operation. The traffic splitting, in turn, is a function of the departure-traffic descriptor as computed by the service operation. An inspection of the involved terms for the traffic merging (Equation (15.1) and (15.2)), the traffic splitting (Equation (15.3)), and the service operation (Equations (15.5), (15.8), and (15.11)) shows that the proof of continuity reduces to the question whether, for a given node, the loss probability π (in case of a finite queue) and the variance σ_0^2 are continuous functions of the arrival traffic $\langle \lambda_a, c_a^2 \rangle$. Since π and σ_0^2 depend on the stationary distribution v of the underlying CTMC we can make use of the following theorem [32] to prove this continuity:

The stationary distribution of a CTMC as function of the transition rates $\lambda_1, \dots, \lambda_n$ *of the generator matrix is continuous for all* $\lambda_i > 0, i = 1, \dots, n$ *if the CTMC has exactly one irreducible set of states.*

The underlying CTMC of a queueing station has exactly one irreducible set of states since it is a QBD. Then, the question is, how do the transition rates of the generator matrix depend on $\langle \lambda_a, c_a^2 \rangle$? FiFiQueues uses the traffic descriptor to determine a PH renewal process that represents the arrival traffic. This arrival PH process is then combined with the service PH process of the node to construct the generator matrix. For fixed c_a^2 the transition rates of the generator matrix are a continuous function of the arrival rate λ_a. The theorem then yields the continuity of v as function of λ_a. However, varying c_a^2 may cause FiFiQueues to change the size and structure of the

PH representation. Such a change also influences the size and structure of the QBD. As a consequence, the theorem can only be applied for values of c_a^2 that do not cause such a change. We obtain:

1. v is continuous for $c_a^2 > 1$, since then the PH distribution always takes the same, hyper-exponential, form.
2. v is continuous for $c_a^2 \in \left(\frac{1}{m+1}, \frac{1}{m}\right)$, for all $m \in \{1, \ldots, 9\}$.

The other cases, i.e., $c_a^2 = \frac{1}{m}, m \in \{1, \ldots, 10\}$, have to be separately discussed. In the following we only show the continuity of π. The proof for σ_0^2 is done in a similar way.

15.9.4 Continuity for $c_a^2 = 1$

We show that

$$\lim_{c_a^2 \nearrow 1} \pi(c_a^2) = \pi(c_a^2 = 1) = \lim_{c_a^2 \searrow 1} \pi(c_a^2)$$

which yields the continuity of π around $c_a^2 = 1$.

15.9.4.1 Case $\mathbf{c_a^2 = 1}$

If $c_a^2 = 1$, the arrival PH distribution is a negative-exponential distribution with rate λ_a. Following the notation used in Section 15.4.5, we obtain the steady-state probability distribution $\mathrm{v}^{=}$ by solving the global balance equations

$$\left.\begin{aligned} \mathrm{v}_0^{=}(-\lambda_a) + \mathrm{v}_1^{=}\mathrm{B}^0 &= 0 \\ \mathrm{v}_0^{=}\lambda_a\beta + \mathrm{v}_1^{=}(-\lambda_a\mathrm{I} + \mathrm{B}) + \mathrm{v}_2^{=}\mathrm{B}^0\beta &= 0 \\ \mathrm{v}_1^{=}\lambda_a\mathrm{I} + \mathrm{v}_2^{=}(-\lambda_a\mathrm{I} + \mathrm{B}) + \mathrm{v}_3^{=}\mathrm{B}^0\beta &= 0 \\ \cdots \\ \mathrm{v}_{\mathrm{K}-1}^{=}\lambda_a\mathrm{I} + \mathrm{v}_{\mathrm{K}}^{=}\mathrm{B} &= 0 \end{aligned}\right\} \tag{15.1}$$

and

$$\mathrm{v}^{=} \cdot 1 = 1,$$

where (β, B) is the service PH process and K is the queueing capacity. The loss probability $\pi^{=}$ is then given as

$$\pi^{=} = \pi(c_a^2 = 1) = \frac{1}{\lambda_a}\mathrm{v}_{\mathrm{K}}^{=}(\lambda_a \otimes \mathrm{I}) \cdot 1 = \mathrm{v}_{\mathrm{K}}^{=} \cdot 1.$$

15.9.4.2 Case $\mathbf{c_a^2 \searrow 1}$

If $c_a^2 > 1$, FiFiQueues selects the PH renewal process (α, A) as representation of the arrival traffic with

$$\mathrm{A} = \begin{pmatrix} -\lambda_0 & 0 \\ 0 & -\lambda_1 \end{pmatrix} \quad \text{and} \quad \alpha = (p, 1-p),$$

where $p = \frac{1}{2} + \frac{1}{2}\sqrt{\frac{c_a^2-1}{c_a^2+1}}$, $\lambda_0 = 2p\lambda_a$ and $\lambda_1 = 2(1-p)\lambda_a$.

Let $\mathrm{v}^>$ be the steady-state probability distribution of the resulting QBD. To ease the following calculations we split the components $\mathrm{v}_\mathrm{i}^>$, i$= 0, \ldots, K$, of the probability distribution vector into two parts $\mathrm{v}_\mathrm{i}^> = (\mathrm{v}_\mathrm{i1}^>, \mathrm{v}_\mathrm{i2}^>)$ where $\mathrm{v}_\mathrm{i1}^>$ and $\mathrm{v}_\mathrm{i2}^>$ are associated with the first resp. the second state of the arrival PH process. The vector $\mathrm{v}^>$ is then determined by the following equations:

$$\mathrm{v}_{01}^>(-\lambda_0) + \mathrm{v}_{11}^>\mathrm{B}^0 = 0, \tag{15.2}$$

$$\mathrm{v}_{02}^>(-\lambda_1) + \mathrm{v}_{12}^>\mathrm{B}^0 = 0, \tag{15.3}$$

$$\mathrm{v}_{01}^> p\lambda_0\beta + \mathrm{v}_{02}^> p\lambda_1\beta + \mathrm{v}_{11}^>(-\lambda_0\mathrm{I} + \mathrm{B}) + \mathrm{v}_{21}^>\mathrm{B}^0\beta = 0, \tag{15.4}$$

$$\mathrm{v}_{01}^>(1-p)\lambda_0\beta + \mathrm{v}_{02}^>(1-p)\lambda_1\beta + \mathrm{v}_{12}^>(-\lambda_1\mathrm{I} + \mathrm{B}) + \mathrm{v}_{22}^>\mathrm{B}^0\beta = 0, \tag{15.5}$$

$$\mathrm{v}_{11}^> p\lambda_0\mathrm{I} + \mathrm{v}_{12}^> p\lambda_1\mathrm{I} + \mathrm{v}_{21}^>(-\lambda_0\mathrm{I} + \mathrm{B}) + \mathrm{v}_{31}^>\mathrm{B}^0\beta = 0, \tag{15.6}$$

$$\mathrm{v}_{11}^>(1-p)\lambda_0\mathrm{I} + \mathrm{v}_{12}^>(1-p)\lambda_1\mathrm{I} + \mathrm{v}_{22}^>(-\lambda_1\mathrm{I} + \mathrm{B}) + \mathrm{v}_{32}^>\mathrm{B}^0\beta = 0, \tag{15.7}$$

$$\cdots$$

$$\mathrm{v}_{\mathrm{K}-1,1}^> p\lambda_0\mathrm{I} + \mathrm{v}_{\mathrm{K}-1,2}^> p\lambda_1\mathrm{I} + \mathrm{v}_{\mathrm{K}1}^>((p-1)\lambda_0\mathrm{I} + \mathrm{B}) + \mathrm{v}_{\mathrm{K}2}^> p\lambda_1\mathrm{I} = 0, \tag{15.8}$$

$$\begin{aligned} &\mathrm{v}_{\mathrm{K}-1,1}^>(1-p)\lambda_0\mathrm{I} + \mathrm{v}_{\mathrm{K}-1,2}^>(1-p)\lambda_1\mathrm{I} + \\ &\quad \mathrm{v}_{\mathrm{K}1}^>(1-p)\lambda_0\mathrm{I} + \mathrm{v}_{\mathrm{K}2}^>(-p\lambda_1\mathrm{I} + \mathrm{B}) = 0, \end{aligned} \tag{15.9}$$

and

$$\mathrm{v}^> \cdot 1 = 1.$$

The loss probability $\pi^>$ of the station is then given as

$$\pi^> = \frac{1}{\lambda_a}\mathrm{v}_\mathrm{K}^>(\mathrm{A}^0 \otimes \mathrm{I}) \cdot 1 = \frac{1}{\lambda_a}(\mathrm{v}_{\mathrm{K}1}^>\lambda_0 + \mathrm{v}_{\mathrm{K}2}^>\lambda_1)\mathrm{I} \cdot 1. \tag{15.10}$$

Summing Equations (15.2) and (15.3), (15.4) and (15.5), …, gives:

$$\left.\begin{aligned} s_0 + t_1\mathrm{B}^0 &= 0 \\ -s_0\beta + s_1\mathrm{I} + t_1\mathrm{B} + t_2\mathrm{B}^0\beta &= 0 \\ -s_1\mathrm{I} + s_2\mathrm{I} + t_2\mathrm{B} + t_3\mathrm{B}^0\beta &= 0 \\ \cdots \\ -s_{K-1}\mathrm{I} + t_K\mathrm{B} &= 0 \end{aligned}\right\} \tag{15.11}$$

where $s_i = \mathrm{v}_\mathrm{i1}^>(-\lambda_0) + \mathrm{v}_\mathrm{i2}^>(-\lambda_1)$ and $t_i = \mathrm{v}_\mathrm{i1}^> + \mathrm{v}_\mathrm{i2}^>$. For $c_a^2 \searrow 1$ we have $p \to \frac{1}{2}$. From this, it follows

$$\lim_{c_a^2 \searrow 1} \lambda_0 = \lim_{c_a^2 \searrow 1} \lambda_1 = \lambda_a,$$

and

$$\lim_{c_a^2 \searrow 1} s_i = -\lambda_a \lim_{c_a^2 \searrow 1} t_i, \qquad i = 0, \ldots, k.$$

By applying these limits to Equation (15.11) we observe their correspondence with Equation (15.1). Hence, we obtain for $c_a^2 \searrow 1$:

$$\lim_{c_a^2 \searrow 1} t_i = \lim_{c_a^2 \searrow 1} (\mathrm{v}_{\mathrm{i1}}^{>} + \mathrm{v}_{\mathrm{i2}}^{>}) = \mathrm{v}_{\mathrm{i}}^{=},$$

which provides, with Equation (15.10), the desired relationship

$$\lim_{c_a^2 \searrow 1} \pi^{>} = \pi^{=}.$$

15.9.4.3 Case $\mathbf{c_a^2 \nearrow 1}$

If $0.5 < c_a^2 < 1$, the arrival PH distribution is a modified hypo-exponential distribution (α, A) as defined in Section 15.4.5 where

$$\mathrm{A} = \begin{pmatrix} -\lambda_0 & \lambda_0 \\ 0 & -\lambda_1 \end{pmatrix} \quad \text{and} \quad \alpha = (1, 0).$$

Let $\mathrm{v}^{<}$ be the steady-state probability distribution of the resulting QBD. Again, we split the components $\mathrm{v}_{\mathrm{i}}^{<}$, i$= 0, \ldots, K$, of the probability distribution vector into two parts $\mathrm{v}_{\mathrm{i}}^{<} = (\mathrm{v}_{\mathrm{i1}}^{<}, \mathrm{v}_{\mathrm{i2}}^{<})$, where $\mathrm{v}_{\mathrm{i1}}^{<}$ and $\mathrm{v}_{\mathrm{i2}}^{<}$ are associated with the first resp. the second state of the arrival PH distribution. The vector $\mathrm{v}^{<}$ is then determined by the following equations:

$$\begin{aligned}
\mathrm{v}_{01}^{<}(-\lambda_0) + \mathrm{v}_{11}^{<}\mathrm{B}^0 &= 0, && (15.12)\\
\mathrm{v}_{01}^{<}\lambda_0 + \mathrm{v}_{02}^{<}(-\lambda_1) + \mathrm{v}_{12}^{<}\mathrm{B}^0 &= 0, && (15.13)\\
\mathrm{v}_{02}^{<}\lambda_1\beta + \mathrm{v}_{11}^{<}(-\lambda_0\mathrm{I} + \mathrm{B}) + \mathrm{v}_{21}^{<}\mathrm{B}^0\beta &= 0, && (15.14)\\
\mathrm{v}_{11}^{<}\lambda_0\mathrm{I} + \mathrm{v}_{12}^{<}(-\lambda_1\mathrm{I} + \mathrm{B}) + \mathrm{v}_{22}^{<}\mathrm{B}^0\beta &= 0, && (15.15)\\
\mathrm{v}_{12}^{<}\lambda_1\mathrm{I} + \mathrm{v}_{21}^{<}(-\lambda_0\mathrm{I} + \mathrm{B}) + \mathrm{v}_{31}^{<}\mathrm{B}^0\beta &= 0, && (15.16)\\
\mathrm{v}_{21}^{<}\lambda_0\mathrm{I} + \mathrm{v}_{22}^{<}(-\lambda_1\mathrm{I} + \mathrm{B}) + \mathrm{v}_{32}^{<}\mathrm{B}^0\beta &= 0, && (15.17)\\
&\ldots\\
\mathrm{v}_{\mathrm{K}-1,2}^{<}\lambda_1\mathrm{I} + \mathrm{v}_{\mathrm{K}1}^{<}(-\lambda_0\mathrm{I} + \mathrm{B}) + \mathrm{v}_{\mathrm{K}2}^{<}\lambda_1\mathrm{I} &= 0, && (15.18)\\
\mathrm{v}_{\mathrm{K}1}^{<}\lambda_0\mathrm{I} + \mathrm{v}_{\mathrm{K}2}^{<}(-\lambda_1\mathrm{I} + \mathrm{B}) &= 0, && (15.19)
\end{aligned}$$

and

$$\mathrm{v}^{<} \cdot \mathbf{1} = 1.$$

The loss probability $\pi^{<}$ of the station is then given as

$$\pi^< = \frac{1}{\lambda_a} \mathrm{v}_{\mathrm{K}}^<(\mathrm{A}^0 \otimes \mathrm{I}) \cdot 1 = \frac{1}{\lambda_a} \mathrm{v}_{\mathrm{K2}}^< \lambda_1 \mathrm{I} \cdot 1. \tag{15.20}$$

From Equation (15.19) we obtain

$$\mathrm{v}_{\mathrm{K2}}^< \lambda_1 \mathrm{I} = \mathrm{v}_{\mathrm{K1}}^< \lambda_0 \mathrm{I} + \mathrm{v}_{\mathrm{K2}}^< \mathrm{B},$$

which yields with Equation (15.20):

$$\pi^< = \frac{1}{\lambda_a} (\mathrm{v}_{\mathrm{K1}}^< \lambda_0 \mathrm{I} + \mathrm{v}_{\mathrm{K2}}^< \mathrm{B}) \cdot 1. \tag{15.21}$$

Using simple substitutions we derive from Equations (15.12)–(15.19):

$$\left.\begin{array}{r} \mathrm{v}_{01}^<(-\lambda_0) + \mathrm{v}_{11}^< \mathrm{B}^0 = 0 \\ (\mathrm{v}_{01}^< \lambda_0 + \mathrm{v}_{12}^< \mathrm{B}^0)\beta + \mathrm{v}_{11}^<(-\lambda_0 \mathrm{I} + \mathrm{B}) + \mathrm{v}_{21}^< \mathrm{B}^0 \beta = 0 \\ (\mathrm{v}_{11}^< \lambda_0 \mathrm{I} + \mathrm{v}_{12}^< \mathrm{B} + \mathrm{v}_{22}^< \mathrm{B}^0 \beta) + \mathrm{v}_{21}^<(-\lambda_0 \mathrm{I} + \mathrm{B}) + \mathrm{v}_{31}^< \mathrm{B}^0 \beta = 0 \\ (\mathrm{v}_{\mathrm{K}-1,1}^< \lambda_0 \mathrm{I} + \mathrm{v}_{\mathrm{K}-1,2}^< \mathrm{B} + \mathrm{v}_{\mathrm{K2}}^< \mathrm{B}^0 \beta) + \mathrm{v}_{\mathrm{K1}}^< \mathrm{B} + \mathrm{v}_{\mathrm{K2}}^< \mathrm{B} = 0 \end{array}\right\} \tag{15.22}$$

For $c_a^2 \nearrow 1$ we have $\lambda_0 \rightarrow \lambda_a$ and $\lambda_1 \rightarrow \infty$. Solving Equations (15.13), (15.15), ..., (15.19) for $\mathrm{v}_{\mathrm{i2}}^<$ then gives

$$\mathrm{v}_{\mathrm{i2}}^< \rightarrow 0, \quad i = 0, \ldots, K.$$

By applying these limits to Equation (15.22) we observe their correspondence with Equation (15.1). Hence we obtain for $c_a^2 \nearrow 1$:

$$\mathrm{v}_{\mathrm{i1}}^< \rightarrow \mathrm{v}_{\mathrm{i}}^= \quad \text{and} \quad \mathrm{v}_{\mathrm{i2}}^< \rightarrow 0, \qquad i = 0, \ldots, K,$$

which gives, applied to Equation (15.21):

$$\lim_{c_a^2 \nearrow 1} \pi^< = \pi^=,$$

as required.

15.9.5 Continuity for $c_a^2 = \frac{1}{m}, m \in \{2, \ldots, 10\}$

The transition rates of the hypo-exponential PH distributions for $c_a^2 < 1$ are defined as functions of the number of phases $m = \left\lceil \frac{1}{c_a^2} \right\rceil$. The inherent discontinuity suggests that the steady-state distribution of the resulting QBD is discontinuous, too. To improve the behavior of the arrival distribution with regard to the continuity, [47] proposes to modify FiFiQueues' PH fitting procedure for $c_a^2 < 1$ as follows. Given the mean inter-arrival time $\mathrm{E}[X] = 1/\lambda_a$ and the squared coefficient of variation c_a^2, we fit the PH distribution (α, A) with $m = \left\lceil \frac{1}{c_a^2} \right\rceil$ phases and initial probability vector

$\alpha = (1,0,\dots,0)$. The matrix A is given as

$$\mathrm{A} = \begin{pmatrix} -\lambda_0 & \lambda_0 & & & & \\ & -\lambda_1 & \lambda_1 & & & \\ & & \ddots & \ddots & & \\ & & & -\lambda_{m-2} & \lambda_{m-2} \\ & & & & -\lambda_{m-1} \end{pmatrix}, \tag{15.23}$$

where

$$\begin{aligned} \lambda_i &= 1/(c_a^2 \mathrm{E}[X]), \text{ for } 0 \le i < m-2, \\ \lambda_{m-2} &= \frac{\lambda_{m-1}}{\mathrm{E}[X]\lambda_{m-1}(1-(m-2)c_a^2) - 1}, \\ \lambda_{m-1} &= \frac{1-(m-2)c_a^2 + \sqrt{(c_a^2)^2(2m-m^2) + 2c_a^2(m-1) - 1}}{\mathrm{E}[X]((m-1)(m-2)(c_a^2)^2 + c_a^2(3-2m) + 1)} \end{aligned}$$

As can be seen, the transition rates λ_i for $i < m-2$ are now continuous functions of c_a^2. This causes that important statistics of this new PH renewal process, e.g., the third moment of the inter-arrival time, are now continuous at values of c_a^2 where a size change happens. Note that this is not true for the original PH distribution. Figure 15.17 illustrates this by comparing the third moment of both PH distributions for values of c_a^2 around 0.5 (i.e., when the size m changes from 3 to 2).

For $c_a^2 \nearrow \frac{1}{m}$, $m \in \{1,\dots,9\}$, the new PH distribution (with $m+1$ phases) yields

$$\lambda_i \longrightarrow m\lambda_a, \qquad \text{for } 0 \le i < m, \tag{15.24}$$

$$\lambda_m \longrightarrow \infty. \tag{15.25}$$

Hence, the proof of $\lim_{c_a^2 \nearrow 1} \pi(c_a^2) = \pi(c_a^2 = 1)$ for $m = 1$, as given in the previous section, is still valid and we only have to prove that

$$\lim_{c_a^2 \nearrow \frac{1}{m}} \pi(c_a^2) = \pi(c_a^2 = \frac{1}{m})$$

for $m \in \{2,\dots,9\}$. For $c_a^2 = \frac{1}{m}$ the arrival PH distribution $(\alpha_=, \mathrm{A}_=)$ is an Erlang-m distribution. Again, let $\mathrm{v}^=$ be the steady-state probability vector of the resulting QBD with $c_a^2 = \frac{1}{m}$, $m \in \{2,\dots,10\}$. We split the components $\mathrm{v}_\mathrm{i}^=$, i$= 0,\dots,K$, of the probability distribution vector into m parts $\mathrm{v}_\mathrm{i}^= = (\mathrm{v}_\mathrm{i1}^=,\dots,\mathrm{v}_\mathrm{im}^=)$ where $\mathrm{v}_\mathrm{ij}^=$ is associated with the j-th state of the arrival PH distribution. The probabilities are determined by $\mathrm{v}^= \cdot \mathbf{1} = 1$ and by the global balance equations of the QBD. For level 0 of the QBD, when no customers are in the queueing station, we obtain

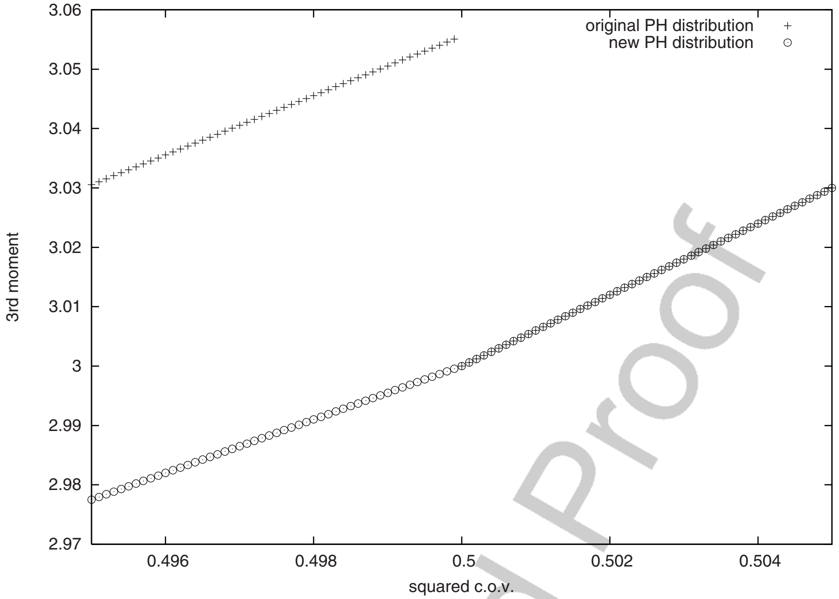


Fig. 15.17: Third moment of the original (Equation (15.4)) and the modified (Equation (15.23)) PH distribution as function of c_a^2

$$
\begin{aligned}
\mathrm{v}^{=}_{01}(-\lambda^{=}) + \mathrm{v}^{=}_{11}\mathrm{B}^0 &= 0,\\
\mathrm{v}^{=}_{01}\lambda^{=} + \mathrm{v}^{=}_{02}(-\lambda^{=}) + \mathrm{v}^{=}_{12}\mathrm{B}^0 &= 0,\\
&\cdots\\
\mathrm{v}^{=}_{0,\mathrm{m}-1}\lambda^{=} + \mathrm{v}^{=}_{0\mathrm{m}}(-\lambda^{=}) + \mathrm{v}^{=}_{1\mathrm{m}}\mathrm{B}^0 &= 0,
\end{aligned}
$$

where $\lambda^{=} = m\lambda_a$. For level $1 \leq i < K$, we have

$$
\begin{aligned}
\mathrm{v}^{=}_{\mathrm{i}-1,\mathrm{m}}\lambda^{=}\beta + \mathrm{v}^{=}_{\mathrm{i}1}(-\lambda^{=}\mathrm{I} + \mathrm{B}) + \mathrm{v}^{=}_{\mathrm{i}+1,1}\mathrm{B}^0\beta &= 0,\\
\mathrm{v}^{=}_{\mathrm{i}1}\lambda^{=}\mathrm{I} + \mathrm{v}^{=}_{\mathrm{i}2}(-\lambda^{=}\mathrm{I} + \mathrm{B}) + \mathrm{v}^{=}_{\mathrm{i}+1,2}\mathrm{B}^0\beta &= 0,\\
&\cdots\\
\mathrm{v}^{=}_{\mathrm{i},\mathrm{m}-1}\lambda^{=}\mathrm{I} + \mathrm{v}^{=}_{\mathrm{im}}(-\lambda^{=}\mathrm{I} + \mathrm{B}) + \mathrm{v}^{=}_{\mathrm{i}+1,\mathrm{m}}\mathrm{B}^0\beta &= 0,
\end{aligned}
$$

and, finally, for level K:

$$
\begin{aligned}
\mathrm{v}^{=}_{\mathrm{K}-1,\mathrm{m}}\lambda^{=}\mathrm{I} + \mathrm{v}^{=}_{\mathrm{K}1}(-\lambda^{=}\mathrm{I} + \mathrm{B}) + \mathrm{v}^{=}_{\mathrm{Km}}\lambda_{m-1}\mathrm{I} &= 0,\\
\mathrm{v}^{=}_{\mathrm{K}1}\lambda^{=}\mathrm{I} + \mathrm{v}^{=}_{\mathrm{K}2}(-\lambda^{=}\mathrm{I} + \mathrm{B}) &= 0,\\
&\cdots\\
\mathrm{v}^{=}_{\mathrm{K},\mathrm{m}-1}\lambda^{=}\mathrm{I} + \mathrm{v}^{=}_{\mathrm{Km}}(-\lambda^{=}\mathrm{I} + \mathrm{B}) &= 0.
\end{aligned}
$$

The loss probability $\pi^{=}$ of the station is given by

$$\pi^{=} = \frac{1}{\lambda_a} \mathrm{v}_{\mathrm{K}}^{=} (\mathrm{A}_{=}^{0} \otimes \mathrm{I}) \cdot 1 = \frac{1}{\lambda_a} \mathrm{v}_{\mathrm{Km}}^{=} \lambda^{=} \mathrm{I} \cdot 1.$$

For $\frac{1}{m+1} \leq c_a^2 < \frac{1}{m}$ the resulting arrival PH process $(\alpha_{<}, \mathrm{A}_{<})$ has $m+1$ states and the steady-state probability vector $\mathrm{v}^{<}$ of the QBD has additional components $\mathrm{v}_{\mathrm{i,m+1}}^{<}, i = 0, \ldots, K$. Let $\lambda_0, \ldots, \lambda_{m+1}$ be the transition rates of the PH process. The global balance equations for level 0 are

$$\begin{aligned} \mathrm{v}_{01}^{<}(-\lambda_0) + \mathrm{v}_{11}^{<}\mathrm{B}^0 &= 0, \\ \mathrm{v}_{01}^{<}\lambda_0 + \mathrm{v}_{02}^{<}(-\lambda_1) + \mathrm{v}_{12}^{<}\mathrm{B}^0 &= 0, \\ \ldots \\ \mathrm{v}_{0\mathrm{m}}^{<}\lambda_{m-1} + \mathrm{v}_{0,\mathrm{m+1}}^{<}(-\lambda_m) + \mathrm{v}_{1,\mathrm{m+1}}^{<}\mathrm{B}^0 &= 0. \end{aligned}$$

For level $1 \leq i < K$, we have

$$\begin{aligned} \mathrm{v}_{\mathrm{i-1,m+1}}^{<}\lambda_m\beta + \mathrm{v}_{\mathrm{i1}}^{<}(-\lambda_0\mathrm{I} + \mathrm{B}) + \mathrm{v}_{\mathrm{i+1,1}}^{<}\mathrm{B}^0\beta &= 0, \\ \mathrm{v}_{\mathrm{i1}}^{<}\lambda_0\mathrm{I} + \mathrm{v}_{\mathrm{i2}}^{<}(-\lambda_1\mathrm{I} + \mathrm{B}) + \mathrm{v}_{\mathrm{i+1,2}}^{<}\mathrm{B}^0\beta &= 0, \\ \ldots \\ \mathrm{v}_{\mathrm{im}}^{<}\lambda_{m-1}\mathrm{I} + \mathrm{v}_{\mathrm{i,m+1}}^{<}(-\lambda_m\mathrm{I} + \mathrm{B}) + \mathrm{v}_{\mathrm{i+1,m+1}}^{<}\mathrm{B}^0\beta &= 0, \end{aligned}$$

and for level K:

$$\mathrm{v}_{\mathrm{K-1,m+1}}^{<}\lambda_m\mathrm{I} + \mathrm{v}_{\mathrm{K1}}^{<}(-\lambda_0\mathrm{I} + \mathrm{B}) + \mathrm{v}_{\mathrm{K,m+1}}^{<}\lambda_m\mathrm{I} = 0 \tag{15.26}$$

$$\mathrm{v}_{\mathrm{K1}}^{<}\lambda_0\mathrm{I} + \mathrm{v}_{\mathrm{K2}}^{<}(-\lambda_1\mathrm{I} + \mathrm{B}) = 0 \tag{15.27}$$

$$\ldots$$

$$\mathrm{v}_{\mathrm{Km}}^{<}\lambda_{m-1}\mathrm{I} + \mathrm{v}_{\mathrm{K,m+1}}^{<}(-\lambda_m\mathrm{I} + \mathrm{B}) = 0. \tag{15.28}$$

The loss probability $\pi^{<}$ is given by

$$\pi^{<} = \frac{1}{\lambda_a}\mathrm{v}_{\mathrm{K}}^{<}(\mathrm{A}_{<}^{0} \otimes \mathrm{I}) \cdot 1 = \frac{1}{\lambda_a}\mathrm{v}_{\mathrm{K,m+1}}^{<}\lambda_m\mathrm{I} \cdot 1. \tag{15.29}$$

From Equation (15.28) we obtain

$$\mathrm{v}_{\mathrm{K,m+1}}^{<}\lambda_m\mathrm{I} = \mathrm{v}_{\mathrm{Km}}^{<}\lambda_{m-1}\mathrm{I} + \mathrm{v}_{\mathrm{K,m+1}}^{<}\mathrm{B},$$

which yields with Equation (15.29):

$$\pi^{<} = \frac{1}{\lambda_a}(\mathrm{v}_{\mathrm{Km}}^{<}\lambda_{m-1}\mathrm{I} + \mathrm{v}_{\mathrm{K,m+1}}^{<}\mathrm{B}) \cdot 1. \tag{15.30}$$

Solving the global balance equations for $\mathrm{v}_{\mathrm{i,m+1}}^{<}$ and applying the limits from (15.24) and (15.25) gives

$$\mathrm{v}_{\mathrm{i,m+1}}^{<} \to 0, \quad i = 0, \ldots, K.$$

Similar to the case $c_a^2 \nearrow 1$ of the original PH distribution (see Equation (15.22)), transforming the global balance equations finally yields

$$\begin{aligned} \mathrm{v}_{\mathrm{ij}}^{<} &\rightarrow \mathrm{v}_{\mathrm{ij}}^{=}, \\ \mathrm{v}_{\mathrm{i,m+1}}^{<} &\rightarrow 0, \qquad\qquad i = 0, \dots, K \text{ and } j = 1, \dots, m, \end{aligned}$$

which gives, applied to Equation (15.30):

$$\lim_{c_a^2 \nearrow \frac{1}{m}} \pi^{<} = \pi^{=},$$

as required.

We now have shown that the loss probability π is a continuous function of c_a^2. Together with the continuity of the variance σ_0^2 (not shown here), this yields the continuity of the the function H, and, as consequence, the existence of the fixed point.

Note that experiments have shown that the original FiFiQueues and the modified version using the new hypo-exponential PH distributions compute nearly identical results with relative errors of less than 10^{-4}.

References

1. Akar, N., Sohraby, K.: An invariant subspace approach in M|G|1 and G|M|1 type Markov chains. Communications in Statistics: Stochastic Models **13(3)**, 251–257 (1997)
2. Apache Software Foundation: Apache HTTP Server Project. http://httpd.apache.org/
3. Bini, D., Chakravarthy, S., Meini, B.: A new algorithm for the design of finite capacity service units. In: Numerical Solution of Markov Chains (NSMC'99), pp. 247–260. Prensas Universitarias de Zaragoza (1999)
4. Bini, D., Meini, B.: On cyclic reduction applied to a class of Toeplitz-like matrices arising in queueing problems. In: Proceedings of the Second International Workshop on Numerical Solution of Markov Chains, pp. 21–38. Raleigh, North Carolina (1995)
5. Bocharov, P.: Analysis of the queue length and the output flow in single server with finite waiting room and phase type distributions. Problems of Control and Information Theory **16**(3), 211–222 (1987)
6. Bocharov, P., Naoumov, V.: Matrix-geometric stationary distribution for the PH/PH/1/r queue. Elektronische Informationsverarbeitung und Kybernetik **22(4)**, 179–186 (1986)
7. Burke, P.: The output of a queueing system. Operations Research **4**, 699–704 (1956)
8. Chakka, R.: Performance and reliability modelling of computing systems using spectral expansion. Ph.D. thesis, University of Newcastle upon Tyne (1995)
9. Chakravarthy, S., Neuts, M.: Algorithms for the design of finite-capacity service units. Naval Research Logistics **36**, 147–165 (1989)
10. Fischer, W., Meier-Hellstern, K.: The Markov-modulated Poisson process (MMPP) cookbook. Performance Evaluation **18**, 149–171 (1992)
11. Gribble, S.: UC Berkeley Home IP HTTP Traces. http://www.acm.org/sigcomm/ITA/

12. Harrison, J., Nguyen, V.: The QNET method for two-moment analysis of open queueing networks. Queueing Systems: Theory and Applications **6**(1), 1–32 (1990)
13. Haverkort, B.: Approximate analysis of networks of PH|PH|1|K queues: Theory & tool support. In: H. Beilner, F. Bause (eds.) MMB, *Lecture Notes in Computer Science*, vol. 977, pp. 239–253. Springer (1995)
14. Haverkort, B.: QNAUT: Approximately analyzing networks of PH|PH|1|K queues. Proceedings of the 1996 International Computer Performance and Dependability Symposium p. 57 (1996)
15. Haverkort, B.: Approximate analysis of networks of PH|PH|1|K queues with customer losses: Test results. Annals of Operations Research **79**, 271–291 (1998)
16. Haverkort, B.: Performance of Computer Communication Systems — A Model-Based Approach. John Wiley & Sons (1998)
17. Heindl, A.: Decomposition of general tandem queueing networks with mmpp input. Performance Evaluation **44**(1-4), 5–23 (2001)
18. Heindl, A.: Traffic-based decomposition of general queueing networks with correlated input processes. Ph.D. thesis, Institut für Technische Informatik, Technische Universität Berlin (2001)
19. Heindl, A.: Decomposition of general queueing networks with MMPP inputs and customer losses. Performance Evaluation **51**(2-4), 117–136 (2003)
20. Heindl, A., Mitchell, K., van de Liefvoort, A.: Correlation bounds for second-order MAPs with application to queueing network decomposition. Performance Evaluation **63**(6), 553 – 577 (2006). Modelling Techniques and Tools for Computer Performance Evaluation
21. Heindl, A., Telek, M.: Output models of MAP/PH/1(/K) queues for an efficient network decomposition. Performance Evaluation **49**(1/4), 321–339 (2002)
22. Jackson, J.: Networks of waiting lines. Operations Research **5**, 518–521 (1957)
23. Johnson, M., Taaffe, M.: Matching moments to phase distributions: Mixtures of Erlang distributions of common order. Communications in Statistics: Stochastic Models **5**(4), 711–743 (1989)
24. Kim, S., Muralidharan, R., O'Cinneide, C.: Taking account of correlations between streams in queueing network approximations. Queueing Systems: Theory and Applications **49**(3–4), 261–281 (2005)
25. Koschel, T.: Modellierung und Bewertung von Verteilten Web-Servern. Diploma thesis, Lehr- und Forschungsgebiet Informatik 4, RWTH Aachen (2002)
26. Krämer, W., Langenbach-Belz, M.: Approximate Formulae for the Delay in the Queueing System GI/G/1. In: Proceedings of the 8th International Teletraffic Congress, pp. 235–1/8 (1976)
27. Kühn, P.: Approximate analysis of general queueing networks by decomposition. IEEE Transactions on Communications **27**(1), 113–126 (1979)
28. Latouche, G., Ramaswami, V.: A logarithmic reduction algorithm for quasi birth and death processes. Journal of Applied Probability **30**, 650–674 (1993)
29. Lucantoni, D.: New results on the single server queue with a batch Markovian arrival process. Commun. Statist.- Stochastic Models **7**(1), 1–46 (1991)
30. Lucantoni, D., Choudhury, G., Whitt, W.: Computing transient distributions in general single-server queues. In: Global Telecommunications Conference (GLOBECOM '93), pp. 1045–1050. IEEE (1993)
31. Lucantoni, D., Meier-Hellstern, K., Neuts, M.: A single server queue with server vacations and a class of non-renewal arrival processes. Advances in Applied Probability **22**, 676–705 (1990)
32. Mainkar, V., Trivedi, K.: Sufficient conditions for existence of a fixed point in stochastic reward net-based iterative models. IEEE Transactions of Software Engineering **22**(9), 640–653 (1996)
33. Marshall, K.: Some inequalities in queueing. Operations Research **16**, 651–665 (1968)
34. Naoumov, V., Krieger, U., Wagner, D.: Analysis of a multi-server delay-loss system with a general Markovian arrival process. Matrix-Analytical Methods in Stochastic Models **183**, 43–66 (1996)

35. Network Working Group: RFC 1945. Hypertext Transfer Protocol – HTTP/1.0. http://www.w3.org/Protocols/rfc1945/rfc1945 (1996)
36. Neuts, M.: A versatile Markovian point process. Journal of Applied Probability **16**(2), 764–779 (1979)
37. Neuts, M.: Matrix-Geometric Solutions in Stochastic Models — An Algorithmic Approach. Dover Publications, Inc. (1994)
38. Ost, A.: Performance of communication systems – a model-based approach with matrix-geometric methods. Ph.D. thesis, Lehr- und Forschungsgebiet Informatik 4, RWTH Aachen (2001)
39. Ruemmler, C., Wilkes, J.: An introduction to disk drive modeling. IEEE Computer **27**(3), 17–29 (1994)
40. Sadre, R.: Decomposition-based analysis of queueing networks. Ph.D. thesis, University of Twente (2006)
41. Sadre, R., Haverkort, B.: FiFiQueues: fixed-point analysis of queueing networks with finite-buffer stations. In: MMB (Kurzvorträge), vol. 99-16, pp. 77–80. Universität Trier (1999)
42. Sadre, R., Haverkort, B.: FiFiQueues: fixed-point analysis of queueing networks with finite-buffer stations. In: Computer Performance Evaluation. Modelling Techniques and Tools: 11th International Conference, TOOLS 2000, *Lecture Notes in Computer Science*, vol. 1786, pp. 324–327. Springer (2000)
43. Sadre, R., Haverkort, B., Ost, A.: An efficient and accurate decomposition method for open finite- and infinite-buffer queueing networks. In: W. Stewart, B. Plateau (eds.) Proc. 3rd Int. Workshop on Numerical Solution of Markov Chains, pp. 1–20. Zaragosa University Press (1999)
44. Sadre, R., Haverkort, B., Reinelt, P.: A fixed-point algorithm for closed queueing networks. In: K. Wolter (ed.) Formal Methods and Stochastic Models for Performance Evaluation, Fourth European Performance Engineering Workshop, EPEW 2007, *Lecture Notes in Computer Science*, vol. 4748, pp. 154–170. Springer (2007)
45. Samelson, H.: On the Brouwer Fixed Point Theorem. Portugaliae mathematica **22**(4), 189–191 (1963)
46. Schuba, M., Haverkort, B., Schneider, G.: Performance evaluation of multicast communication in packet-switched networks. Performance Evaluation **39**(1-4), 61–80 (2000)
47. Tartemann, D.: Untersuchung der Existenz eines Fixpunktes in einem iterativen Verfahren zur Warteschlangenanalyse. Diploma thesis, Lehr- und Forschungsgebiet Informatik 4, RWTH Aachen (2002)
48. Weerstra, A.: Using matrix-geometric methods to enhance the QNA method for solving large queueing networks. Diploma thesis, Department of Computer Science, University of Twente (1994)
49. Whitt, W.: The Queueing Network Analyzer. The Bell System Technical Journal **62**(9), 2779–2815 (1983)
50. Whitt, W.: Performance of The Queueing Network Analyzer. The Bell System Technical Journal **62**(9), 2817–2843 (1983)

Chapter 16
Loss Networks

Stan Zachary and Ilze Ziedins

Abstract This chapter reviews the theory of loss networks, in which calls of various types are accepted for service provided that this can commence immediately; otherwise they are rejected. An accepted call remains in the network for some holding time, which is generally independent of the state of the network, and throughout this time requires capacity simultaneously from various network resources. Both equilibrium and dynamical behaviour are studied; for the former a new approach is taken to the theory of uncontrolled loss networks, while the latter is the key to the understanding of stability issues in such networks.

16.1 Introduction

In a loss network calls, or customers, of various types are accepted for service provided that this can commence immediately; otherwise they are rejected. An accepted call remains in the network for some holding time, which is generally independent of the state of the network, and throughout this time requires capacity simultaneously from various network resources.

The loss model was first introduced by Erlang as a model for the behaviour of just a single telephone link (see Brockmeyer *et al.* [7]). The typical example remains that of a communications network, in which the resources correspond to the links in the network, and a call of any type requires, for the duration of its holding time, a fixed allocation of capacity from each link over which it is routed (Kelly [26]). This is the case for a traditional circuit-switched telephone network, but the model is also

Stan Zachary
Heriot-Watt University, Edinburgh, UK
e-mail: s.zachary@hw.ac.uk

Ilze Ziedins
University of Auckland, Auckland, New Zealand
e-mail: i.ziedins@auckland.ac.nz

R.J. Boucherie and N.M. van Dijk (eds.), *Queueing Networks: A Fundamental Approach*,
International Series in Operations Research & Management Science 154,
DOI 10.1007/978-1-4419-6472-4_16, © Springer Science+Business Media, LLC 2011

appropriate to modern computer communications networks which support streaming applications with minimum bandwidth requirements (Kelly *et al.* [28]). There are also other examples: for instance, in a cellular mobile network similar capacity constraints arise from the need to avoid interference (Abdalla and Boucherie [1]).

The mathematics of such networks has been widely studied, with interest in both equilibrium and, more recently, dynamical behaviour. Of particular importance are questions of call acceptance and capacity allocation (for example, routing), with the aim of ensuring good network performance which is additionally robust with respect to variations in network parameters. Call arrival rates, in particular, may fluctuate greatly. An excellent review of the state-of-the-art at the time of its publication is given by Kelly [27]—see also the many papers cited therein, and the later survey by Ross [39].

We take as our model the following. Let $\mathcal{R}$ denote the finite set of possible call, or customer, types. Calls of each type $r \in \mathcal{R}$ arrive at the network as a Poisson process with rate ν_r, and each such call, if accepted by the network (see below), remains in it for a *holding time* which is exponentially distributed with mean μ_r^{-1}. We shall discuss later the extent to which these assumptions, in particular the latter, are necessary. Calls which are rejected do not retry and are simply considered lost. All arrival processes and holding times are independent of one another. We denote the state of the network at time t by $\mathbf{n}(t) = (n_r(t), r \in \mathcal{R})$, where $n_r(t)$ is the number of calls of each type r in progress at that time. The process $\mathbf{n}(\cdot)$ is thus Markov. It takes values in some state space $\mathcal{N} \subset \mathbb{Z}_+^R$, where $R = |\mathcal{R}|$. We assume $\mathcal{N}$ to be defined by a number of resource constraints

$$\sum_{r \in \mathcal{R}} A_{jr} n_r \le C_j, \qquad j \in \mathcal{J}, \tag{16.1}$$

indexed in a finite set $\mathcal{J}$, where the A_{jr} and the C_j are nonnegative integers. Typically we think of a call of each type r as having a simultaneous requirement, for the duration of its holding time, for A_{jr} units of the capacity C_j of each resource j; however, we show below that the resource constraints (16.1) can also arise in other ways. As noted above, in applications of this model to communications networks, the network resources usually correspond to the *links* in the network, and when discussing the model in that context we shall generally find it convenient to use this terminology. We shall also find it helpful to define, for each $r \in \mathcal{R}$, the parameter $\kappa_r = \nu_r/\mu_r$; many quantities of interest depend on ν_r and μ_r only through their ratio κ_r.

We shall say that a network is *uncontrolled* whenever calls are accepted subject only to the condition that the resulting state of the network belongs to the set $\mathcal{N}$. Uncontrolled networks are particularly amenable to mathematical analysis and are in certain senses very well-behaved. In addition, such a network has the important *insensitivity* property: the stationary distribution of the process $\mathbf{n}(\cdot)$ is unaffected by the relaxation of the assumption that the call holding time distributions are exponential, and depends on these holding time distributions only through their means. This is essentially a consequence of the *detailed balance* property considered in Section 16.2.1.

However, as we shall also see, the performance of uncontrolled networks may be far from optimal. A more general control strategy is given by requiring that a call of type r, which arrives when the state of the network (immediately prior to its arrival) is $\mathbf{n}$, is accepted if and only if $\mathbf{n} \in \mathcal{A}_r$ for some *acceptance set* $\mathcal{A}_r$. The sets $\mathcal{A}_r$ may be chosen so as to optimise, in some appropriate sense, the network's performance. Such networks do not in general possess the insensitivity property described above.

Of interest in a loss network are both the stationary distribution π and the dynamics of the process $\mathbf{n}(\cdot)$. For the former it is usual to compute, for each r, the stationary *blocking probability* B_r, that a call of type r is rejected; here we shall find it slightly easier to work with the stationary *acceptance* (or *passing*) *probability* $P_r = 1 - B_r$. We note immediately that, by Little's Theorem, the stationary expected number of calls of each type r in the network is given by

$$\mathbf{E}_\pi n_r = \kappa_r P_r, \tag{16.2}$$

where κ_r is as defined above. Thus acceptance probabilities may be regarded as one of the key performance measures in the stationary regime.

It will be convenient to refer to the above model of a loss network—in which arriving calls have fixed resource requirements and in which the only control in the network is the ability to reject calls—as the *canonical model*. When considering communications networks, it is natural to extend this model by allowing also the possibility of *alternative*, or *state-dependent*, *routing*, in which calls choose their route according to the current state of the network. Here the state space should properly be expanded to record the number of calls of each type on each route (but see below). We consider such models in Section 16.3.2.

In the case where we allow not only alternative routing, but also *repacking* of calls already in the network, the model simplifies again, and it is once more only necessary for the state space to record the number of calls of each type in progress. Consider the simple example of a communications network consisting of three links with capacities $\hat{C}_1, \hat{C}_2, \hat{C}_3$, and three call types, in which calls of each type $r = 1,2,3$ require *either* one unit of capacity from the corresponding link r *or* one unit of capacity from each of the other two links (in each case the distribution of the call holding time is assumed to be the same). If repacking is allowed, the state of the system may be given by $\mathbf{n} = (n_1, n_2, n_3)$ as usual, and it is easy to check that a call of any type may be accepted if and only if the resulting state of the network satisfies the constraints

$$n_r + n_{r'} \le \hat{C}_r + \hat{C}_{r'}, \qquad r \ne r', \qquad r, r' \in \mathcal{R}.$$

This is therefore an instance of the uncontrolled network discussed above, in which the coefficients A_{jr} and C_j must be appropriately defined.

Exact calculations for large loss networks typically exceed the capabilities of even large computers, and we are thus led to consider approximations. Mathematical justification for these approximations is usually based on asymptotic results for one of two limiting schemes. In the first, which we shall refer to as the *Kelly limiting scheme* (see Kelly [26]), the sets $\mathcal{R}$, $\mathcal{J}$, the matrix $A = (A_{jr})$, and the parameters μ_r

are held fixed, while the arrival rates ν_r and the capacities C_j are all allowed to increase in proportion to a *scale parameter N* which tends to infinity. In the second, which is known as the *diverse routing limit* (see Whitt [42], and Ziedins and Kelly [47]), the capacity of each resource is held constant, while the sets $\mathcal{R}$ and $\mathcal{J}$ (and correspondingly the size of the matrix A) are allowed to increase, and the arrival rates for call types requiring capacity at more than one resource to decrease, in such a way that the total traffic offered to each resource is also held constant (in particular, this requires that the arrival rate for any call type that requires capacity at more than one resource becomes negligible in the limit). Results for the latter scheme in particular are used to justify assumptions of independence in many approximations.

For each time t, define $\mathbf{m}(t) = (m_j(t), j \in \mathcal{J})$, where $m_j(t)$ denotes the current occupancy, or usage, of each resource j in a loss network. Define also π' to be the stationary distribution of the process $\mathbf{m}(\cdot)$. In particular, for the canonical model defined above, for each t,

$$m_j(t) = \sum_{r \in \mathcal{R}} A_{jr} n_r(t); \tag{16.3}$$

here the process $\mathbf{m}(\cdot)$ takes values in the set

$$\mathcal{M} = \{\mathbf{m} \in \mathbb{Z}_+^J : 0 \le m_j \le C_j,\ j \in \mathcal{J}\}, \tag{16.4}$$

where we write $J = |\mathcal{J}|$, and the distribution π' is given by

$$\pi'(\mathbf{m}) = \sum_{\mathbf{n}:\, A\mathbf{n}=\mathbf{m}} \pi(\mathbf{n}), \qquad \mathbf{m} \in \mathcal{M}. \tag{16.5}$$

In general the process $\mathbf{m}(\cdot)$ takes values in a space of significantly lower dimension than that of the process $\mathbf{n}(\cdot)$. This is especially so in models of communications networks which incorporate alternative routing. It is a recurrent theme in the study of loss networks that, in general, at least approximately optimal control of a network is obtained by basing admission decisions and, in communications networks, routing decisions, solely on the state of the process $\mathbf{m}(\cdot)$ at the arrival time of each call. Further, in this case, a knowledge of the distribution π' is sufficient to determine call acceptance probabilities. We shall also see that good estimates of π' are generally given by assuming its (approximate) factorisation as

$$\pi'(\mathbf{m}) = \prod_{j \in \mathcal{J}} \pi'_j(m_j), \tag{16.6}$$

where each π'_j is normalised to be a probability distribution. This is a further recurrent theme in the study of loss networks.

In Section 16.2 we consider the stationary behaviour of uncontrolled networks, reviewing both exact results and approximations for large networks. Our approach is based on the use of an elegant recursion due to Kaufman [25] and to Dziong and Roberts [12] which delivers all the classical results in regard to, for example, stationary acceptance probabilities, with a certain simplicity.

More general networks are studied in Section 16.3. In Section 16.3.1 we study the problem of optimal control in a single-resource network, where a reasonably tractable analysis of stationary behaviour is again possible, and where we show that either exactly or approximately optimal control may be obtained with the use of strategies based on *reservation* parameters. In Section 16.3.2 we consider multiple-resource networks, allowing in particular the possibility of alternative routing. We again derive approximations which are known to work extremely well in practice. In Section 16.4 we consider the dynamical behaviour of large loss networks. This is important for the study of the long-run, and hence also the equilibrium, behaviour of networks in the case where a direct equilibrium analysis is impossible. The study of network dynamics is also the key to understanding their stability. Finally, in Section 16.5 we mention some wider models and discuss some open problems.

16.2 Uncontrolled loss networks: stationary behaviour

We study here the stationary behaviour of the uncontrolled network introduced above, in which calls of any type are accepted subject only to the condition that the resulting state $\mathbf{n}$ of the network belongs to the state space $\mathcal{N}$ defined by the capacity constraints (16.1). In particular we shall see, in Section 16.2.3 and subsequently, that most quantities of interest, in particular acceptance probabilities, may be calculated, exactly or approximately, without the need to calculate the full stationary distribution π of the process $\mathbf{n}(\cdot)$.

16.2.1 The stationary distribution

For each $r \in \mathcal{R}$, let δ_r be the vector whose rth component is 1 and whose other components are 0. Recall that, under the assumptions introduced above, $\mathbf{n}(\cdot)$ is a Markov process. For $\mathbf{n}, \mathbf{n} - \delta_r \in \mathcal{N}$ and $r \in \mathcal{R}$, its transition rates between $\mathbf{n}$ and $\mathbf{n} - \delta_r$ are $n_r\mu_r$ and ν_r. It thus follows that the stationary distribution π of the process $\mathbf{n}(\cdot)$ is given by the solution of the *detailed balance equations*

$$\pi(\mathbf{n})n_r\mu_r = \pi(\mathbf{n} - \delta_r)\nu_r, \qquad r \in \mathcal{R}, \quad \mathbf{n} \in \mathcal{N}, \tag{16.1}$$

where, here and elsewhere, we make the obvious convention that $\pi(\mathbf{n} - \delta_r) = 0$ whenever $n_r = 0$. That is,

$$\pi(\mathbf{n}) = G^{-1} \prod_{r \in \mathcal{R}} \frac{\kappa_r^{n_r}}{n_r!}, \qquad \mathbf{n} \in \mathcal{N}, \tag{16.2}$$

where the normalising constant G^{-1} is determined by the requirement $\sum_{\mathbf{n} \in \mathcal{N}} \pi(\mathbf{n}) = 1$. The simple *product form* of the stationary distribution (16.2) is a consequence of the fact that the equations (16.1) *do* have a solution, that is, it is a consequence of the reversibility of the stationary version of the process $\mathbf{n}(\cdot)$. Note also that here the stationary distribution π depends on the parameters ν_r and μ_r only through their ratios $\kappa_r = \nu_r/\mu_r$, $r \in \mathcal{R}$. This result is not in general true for networks with controls.

In the variation of our model in which calls of each type r have holding times which are no longer necessarily exponential (but with unchanged mean μ_r^{-1}), it is well-known that the stationary distribution π of the process $\mathbf{n}(\cdot)$ continues to satisfy the detailed balance equations (16.1) and hence also (16.2). For proofs of this *insensitivity* property, see Burman *et al.* [8], Pechinkin [36], or Zachary [44].

The stationary probability that a call of type r is *accepted*, is given by

$$P_r = \sum_{\mathbf{n} \in \mathcal{N}_r} \pi(\mathbf{n}), \tag{16.3}$$

where $\mathcal{N}_r = \{\mathbf{n} \in \mathcal{N} \colon \mathbf{n} + \delta_r \in \mathcal{N}\}$. In Section 16.2.3 we give a recursion which permits a reasonably efficient calculation of the probabilities P_r in networks of small to moderate size. However, the exact calculation of acceptance probabilities is usually difficult or impossible in large networks. We shall therefore also discuss various approximations.

16.2.2 The single resource case

Consider first the case $\mathcal{R} = \{1\}$ of a single call type. For convenience we drop unnecessary subscripts denoting dependence on $r \in \mathcal{R}$; in particular we write $\kappa = \nu/\mu$. We then have $\mathcal{N} = \{n \colon n \leq C\}$ for some positive integer C. The stationary distribution π is a truncated Poisson distribution, and the stationary acceptance probability P is given by Erlang's well-known formula, that is, by $P = 1 - \pi(C) = 1 - E(\kappa, C)$, where

$$E(\kappa, C) = \frac{\kappa^C/C!}{\sum_{n=0}^{C} \kappa^n/n!}. \tag{16.4}$$

Note also that, from (16.2), the expected number of calls in progress under the stationary distribution π is given by κP.

While exact calculation of blocking probabilities via Erlang's formula (16.4) is straightforward, it nevertheless provides insight to give approximations for networks in which C and κ are both large. Formally, we consider the Kelly limiting scheme in which C and κ are allowed to tend to infinity in proportion to a scale parameter N with $p = C/\kappa$ held fixed. The cases $p > 1$, $p = 1$ and $p < 1$ correspond to the network being, in an obvious sense, underloaded, critically loaded, and overloaded respectively. A relatively straightforward analysis of (16.4) shows that,

$$P \to \min(1, p) \qquad \text{as } N \to \infty. \tag{16.5}$$

For $p \geq 1$ the error in the approximation $P \approx 1$ may be estimated by replacing the truncated Poisson distribution of n by a truncated normal distribution. For $p > 1$ the error is thus given asymptotically by $\kappa^{-1/2}\varphi(\kappa^{-1/2}(C-\kappa))$, where φ is the standard normal density function; this decays at least exponentially fast in N. For the critically loaded case $p = 1$ the error is given asymptotically by $(2/\pi\kappa)^{1/2}$ which is $O(N^{-1/2})$ as $N \to \infty$. For the overloaded case $p < 1$ the approximation $P \approx p$ may be refined as follows. Observe that in this case, and since κ and C are large, it follows from either (16.1) or (16.2) that the stationary distribution of *free capacity* in the network is approximately geometric and so the stationary expected free capacity is given by the approximation

$$C - \mathbf{E}_\pi n \approx \frac{p}{1-p}. \tag{16.6}$$

Combining this with (16.2) leads to the very much more refined approximation for the stationary acceptance probability given by

$$P \approx p - \frac{p}{\kappa(1-p)}. \tag{16.7}$$

It thus follows that the error in the original approximation $P \approx p$ is $O(N^{-1})$ as $N \to \infty$.

16.2.3 The Kaufman-Dziong-Roberts (KDR) recursion

For the general model of an uncontrolled network, we now take the set $\mathcal{N}$ to be given by a set of capacity constraints of the form (16.1). We give here an efficient recursion for the determination of stationary acceptance probabilities, due in the case $\mathcal{J} = \{1\}$ to Kaufman [25] and in the general case to Dziong and Roberts [12].

Recall that π' is the stationary distribution of the process $\mathbf{m}(\cdot)$ defined in the Introduction. Since a call of type r arriving at time t is accepted if and only if $m_j(t-) + A_{jr} \leq C_j$ for all j such that $A_{jr} \geq 1$ (where $\mathbf{m}(t-)$ denotes the state of the process $\mathbf{m}(\cdot)$ immediately prior to the arrival of the call), it follows that a knowledge of π' is sufficient to determine stationary acceptance probabilities. Typically the size J of the set $\mathcal{J}$ is smaller than the size R of the set $\mathcal{R}$, and so the dimension of the space $\mathcal{M}$ defined by (16.4) is smaller than that of $\mathcal{N}$. Thus a direct calculation of π', avoiding that of π, is usually much more efficient for determining acceptance probabilities.

For each $r \in \mathcal{R}$, define the vector $\mathbf{A}_r = (A_{jr},\, j \in \mathcal{J})$. For each $\mathbf{m} \in \mathcal{M}$ and $r \in \mathcal{R}$, summing the detailed balance equations (16.1) over $\mathbf{n}$ such that $A\mathbf{n} = \mathbf{m}$ and using also (16.5) yields

$$\kappa_r \pi'(\mathbf{m} - \mathbf{A}_r) = \mathbf{E}(n_r \mid \mathbf{m})\pi'(\mathbf{m}), \qquad r \in \mathcal{R}, \quad \mathbf{m} \in \mathcal{M}, \tag{16.8}$$

where

$$\mathbf{E}(n_r \mid \mathbf{m}) = \frac{\sum_{\mathbf{n}:\, A\mathbf{n}=\mathbf{m}} n_r \pi(\mathbf{n})}{\sum_{\mathbf{n}:\, A\mathbf{n}=\mathbf{m}} \pi(\mathbf{n})}$$

is the stationary expected value of n_r given $A\mathbf{n} = \mathbf{m}$. Since, for each $\mathbf{m}$ and each j, we have $\sum_{r\in\mathcal{R}} A_{jr}\mathbf{E}(n_r \mid \mathbf{m}) = m_j$, it follows from (16.8) that

$$\sum_{r\in\mathcal{R}} A_{jr}\kappa_r\pi'(\mathbf{m}-\mathbf{A}_r) = m_j\pi'(\mathbf{m}), \qquad \mathbf{m}\in\mathcal{M}, \quad j\in\mathcal{J}. \tag{16.9}$$

This is the Kaufman-Dziong-Roberts (KDR) recursion on the set $\mathcal{M}$, enabling the direct determination of successive values of $\pi'(\mathbf{m})$ as multiples of $\pi'(\mathbf{0})$. The entire distribution π' is then determined uniquely by the requirement that $\sum_{\mathbf{m}\in\mathcal{M}} \pi'(\mathbf{m}) = 1$.

16.2.4 Approximations for large networks

We now suppose that κ_r, $r\in\mathcal{R}$, and C_j, $j\in\mathcal{J}$, are sufficiently large that the exact calculation of the stationary distributions π or π' is impracticable. Various bounds for the corresponding stationary acceptance probabilities P_r are given by Whitt [42], who shows in particular that, in the case where $A_{jr}\in\{0,1\}$ for all j and for all r,

$$P_r \geq \prod_{j\in\mathcal{J}} \left(1 - E\left(\sum_{r\in\mathcal{R}} A_{jr}\kappa_r, C_j\right)\right)^{A_{jr}}, \qquad r\in\mathcal{R}, \tag{16.10}$$

where E is the Erlang function (16.4). However, while this bound is intuitively unsurprising, the right side of (16.10) does not provide a very satisfactory approximation for each P_r, as it fails to take account of the fact that the total load at each resource j is effectively reduced by blocking at the remaining resources. We now seek good approximations—for general A_{jr}—for both the stationary distributions π or π' and the corresponding acceptance probabilities P_r, where in all cases account is at least implicitly taken of the "reduced load" phenomenon. We discuss the analytical accuracy of these various approximations. Numerical investigations are performed in the papers cited below.

A simple approximation

We give first a simple approximation, due to Kelly [26], which generalises the approximation $P\approx\min(1,p)$ of Section 16.2.2 for the single-resource case. To provide asymptotic justification we again consider the Kelly limiting scheme, in which the parameters κ_r and C_j are allowed to increase in proportion to a scale parameter N, the sets $\mathcal{R}$, $\mathcal{J}$ and the matrix A being held fixed. We assume (this is largely for simplicity) that the matrix A is such that for each $\mathbf{m}\in\mathcal{M}$ there is at least one $\mathbf{n}\in\mathbb{Z}$ such that $A\mathbf{n} = \mathbf{m}$. (This implies in particular that the matrix A is of full rank.) We outline an argument based on the equations (16.8) and the KDR recursion (16.9).

Suppose that $\pi'(\mathbf{m})$ is maximised at $\mathbf{m}^* \in \mathcal{M}$. The distribution (16.2) of π is a truncation of a product of independent Poisson distributions each of which has a standard deviation which is $O(N^{1/2})$ as the scale parameter N increases. From this and from the mapping of π to π', it follows that all but an arbitrarily small fraction of the distribution of π' is concentrated within a region $\mathcal{M}^* \subseteq \mathcal{M}$ such that the components of $\mathbf{m} \in \mathcal{M}^*$ differ from those of $\mathbf{m}^*$ by an amount which is again $O(N^{1/2})$ as N increases. Further, it is not too difficult to show from the above condition on the matrix A that, for each r, $\mathbf{E}(n_r \mid \mathbf{m})$ varies smoothly with $\mathbf{m}$, in particular in the sense that within $\mathcal{M}^*$ we may make the approximation $\mathbf{E}(n_r \mid \mathbf{m}) \approx \mathbf{E}(n_r \mid \mathbf{m}^*)$, the error yet again being $O(N^{1/2})$ as N increases. It now follows from (16.8) that within $\mathcal{M}^*$ we have

$$\pi'(\mathbf{m}) \approx \pi'(\mathbf{m}^*) \prod_{j\in\mathcal{J}} p_j^{m_j^* - m_j}, \tag{16.11}$$

where necessarily, since $\mathbf{m}^*$ maximises $\pi'(\mathbf{m})$,

$$0 \le p_j \le 1, \qquad j \in \mathcal{J}, \tag{16.12}$$
$$p_j = 1, \qquad \text{for } j \text{ such that } m_j^* < C_j\,. \tag{16.13}$$

Further, from (16.9),

$$\sum_{r\in\mathcal{R}} A_{jr}\kappa_r \prod_{k\in\mathcal{J}} p_k^{A_{kr}} = m_j^* \le C_j, \qquad j \in \mathcal{J}. \tag{16.14}$$

Thus, from (16.11), within $\mathcal{M}^*$ the stationary distribution π' of $\mathbf{m}(\cdot)$ does indeed have the approximate factorisation (16.6), where each of the component distributions π'_j is here geometric (and where in the case $p_j = 1$ the geometric distribution becomes uniform). Further, for each r and each j, we have

$$\pi'_j(\{m_j\colon m_j \le C_j - A_{jr}\}) \approx p_j^{A_{jr}}.$$

Thus the stationary acceptance probabilities P_r are given by the approximation

$$P_r \approx \prod_{j\in\mathcal{J}} p_j^{A_{jr}}, \qquad r \in \mathcal{R}. \tag{16.15}$$

Kelly [26] considered an optimisation problem from which it follows that the equations (16.12)–(16.14) determine the vectors $\mathbf{m}^*$ and $\mathbf{p} = (p_j, j \in \mathcal{J})$ uniquely. He further showed, in an approach based on consideration of the stationary distribution π, that the approximation (16.15) becomes exact as the scale parameter N tends to infinity.

A refined approximation

The *(multiservice) reduced load* or *knapsack approximation* (Dziong and Roberts [12], see also Ross [39]) is a more refined approximation than that defined above. It is given by retaining the approximate factorisation (16.6) of the stationary distribution π' of $\mathbf{m}(\cdot)$. However, subject to this assumed factorisation, the estimation of the component distributions π'_j is refined.

For each $j \in \mathcal{J}$ and $r \in \mathcal{R}$, define

$$p_{jr} = \sum_{m_j=0}^{C_j - A_{jr}} \pi'_j(m_j); \tag{16.16}$$

note that $p_{jr} = 1$ if $A_{jr} = 0$. For fixed j, substitution of (16.6) into the KDR recursion (16.9) and summation over all m_k for all $k \neq j$ yields

$$\sum_{r \in \mathcal{R}} A_{jr} \left(\kappa_r \prod_{k \neq j} p_{kr} \right) \pi'_j(m_j - A_{jr}) = m_j \pi'_j(m_j), \qquad 1 \leq m_j \leq C_j, \quad j \in \mathcal{J} \tag{16.17}$$

(where, as usual, we make the convention $\pi'_j(m_j) = 0$ for $m_j < 0$). This is the one-dimensional KDR recursion associated with a single resource constraint j, and is readily solved to determine π'_j and hence the probabilities p_{jr}, $r \in \mathcal{R}$, in terms of the probabilities p_{kr}, $r \in \mathcal{R}$, for all $k \neq j$. We are thus led to a set of fixed point equations in the probabilities p_{jr}, for which the existence—but not always the uniqueness, see Chung and Ross [10]—of a solution is guaranteed. From (16.6), the probability that a call of type r is accepted is then given by

$$P_r = \prod_{j \in \mathcal{J}} p_{jr}. \tag{16.18}$$

We remark that the recursion (16.17) corresponds to a modified network in which there is a single resource constraint j and each arrival rate κ_r is reduced to $\kappa_r \prod_{k \neq j} p_{kr}$. This *reduced load approximation* is of course exact in the case of a single-resource network.

In the case where each A_{jr} can only take the values 0 or 1 we may set $p_j = p_{jr}$ for r such that $A_{jr} = 1$. The fixed point equations (16.16) and (16.17) then reduce to

$$p_j = 1 - E\left(\sum_{r \in \mathcal{R}} \kappa_r \prod_{k \neq j} p_k^{A_{kr}}, C_j \right) \tag{16.19}$$

where E is again the Erlang function (16.4). This case is the well-known *Erlang fixed point approximation* (EFPA) and has a unique solution, see Kelly [26], and also Ross [39]. It yields acceptance probabilities which are known to be asymptotically exact in the Kelly limiting scheme discussed above, and also, under appropriate conditions, in the *diverse routing* limit discussed in the Introduction—see Whitt [42], and Ziedins and Kelly [47]. The EFPA also has an extension to the case of general A_{jr}, which may be regarded as a simplified version of the reduced load approximation. As with the latter approximation the EFPA may here have multiple solutions.

16.3 Controlled loss networks: stationary behaviour

We now study the more general version of a loss network, in which calls are subject to acceptance controls, and the issues are those of achieving optimal performance.

16.3.1 Single resource networks

We consider a simple model which illustrates some ideas of optimal control—in particular those of *robustness* of the control strategy with respect to variations in arrival rates (which may in practice be unknown, or vary over time).

Suppose that $\mathcal{R} = \{1,2\}$ and that as usual calls of each type r arrive at rate ν_r and have holding times which are exponentially distributed with mean μ_r^{-1}. Suppose further that there is a single resource of capacity C and that a call of either type requires one unit of this capacity, so that the constraints (16.1) here reduce to $n_1 + n_2 \leq C$. We assume that calls of type 1 have greater value per unit time than those of type 2, so that it is desirable to choose the acceptance regions $\mathcal{A}_r$, $r = 1,2$, so as to maximise the linear function

$$\phi(P_1,P_2) := a_1\kappa_1 P_1 + a_2\kappa_2 P_2, \tag{16.1}$$

for some $a_1 > a_2 > 0$ (where, again as usual, for each call type r, $\kappa_r = \nu_r/\mu_r$ and P_r is the stationary acceptance probability.) An upper bound for the expression in (16.1) is given by the solution of the linear programming problem, in the variables P_1, P_2,

$$\text{maximise } \phi(P_1,P_2), \quad \text{subject to } P_r \in [0,1] \text{ for } r = 1,2, \quad \kappa_1 P_1 + \kappa_2 P_2 \leq C \tag{16.2}$$

(where the latter constraint follows from (16.2)). It is easy to see that the solution of this problem is characterised uniquely by the conditions

$$P_1 = P_2 = 1, \quad \text{whenever } \kappa_1 P_1 + \kappa_2 P_2 < C, \tag{16.3}$$

$$P_2 = 0, \qquad \text{whenever } P_1 < 1. \tag{16.4}$$

It is clearly not possible to choose the acceptance regions $\mathcal{A}_1$, $\mathcal{A}_2$ so that the corresponding values of P_1, P_2 solve exactly the problem (16.2). However, we show below that this solution may be achieved asymptotically as the size of the system is allowed to increase, and further that there is an asymptotically optimal control that is both simple and robust with respect to variations in the parameters κ_1, κ_2.

We consider first the form of the optimal control in the special case $\mu_1 = \mu_2$. Here since, at the arrival time t of any call, those calls already within the system are indistinguishable with respect to type, it is clear that the optimal decision on call admission is a function only of the arriving call type and of the total volume $m(t-) = n_1(t-)+n_2(t-)$ of calls already in the system. A formal proof is a straightforward exercise in Markov decision theory. Further, simple coupling arguments show that, for an incoming call of either type arriving at time t and any $0 < m < C$, if it is advantageous to accept the call when $m(t-) = m$, then it is also advantageous to accept the call when $m(t-) = m-1$. It follows that the optimal acceptance regions are of the form

$$\mathcal{A}_1 = \{\mathbf{n}\colon n_1+n_2 < C\} \tag{16.5}$$

$$\mathcal{A}_2 = \{\mathbf{n}\colon n_1+n_2 < C-k\} \tag{16.6}$$

for some *reservation parameter* k, whose optimal value depends on C, κ_1 and κ_2.

Consider now the general case where we do not necessarily have $\mu_1 = \mu_2$, and suppose that C, ν_1 and ν_2 are large. More formally we again have in mind the Kelly limiting scheme in which these parameters are allowed to increase in proportion to some scale parameter N which tends to infinity (while μ_1, μ_2 are held fixed). We further suppose that the acceptance regions are again as given by (16.5) and (16.6), where the reservation parameter k increases slowly with N, i.e. in such a way that

$$k \to \infty, \qquad k/C \to 0, \qquad \text{as } N \to \infty. \tag{16.7}$$

It is convenient to let P_1, P_2 denote the limiting acceptance probabilities. In the case $\kappa_1 + \kappa_2 \leq C$, it is not difficult to see that, since $k/C \to 0$ as $N \to \infty$, we have $P_1 = P_2 = 1$, so that P_1, P_2 solve the optimisation problem (16.2). Consider now the case $\kappa_1 + \kappa_2 > C$. Here, again since $k/C \to 0$ as $N \to \infty$, it follows that, in the limit, the capacity of the network is fully utilised. Further, if κ_1 is sufficiently large that $P_1 < 1$ (informally, even for large N, calls of type 1 are being rejected in significant numbers), then the effect of the increasing reservation parameter k is such that, again in the limit, the network remains sufficiently close to capacity to ensure that no calls of type 2 are accepted, and hence $P_2 = 0$. It now follows that when $\kappa_1 + \kappa_2 > C$, the limiting acceptance probabilities P_1, P_2 satisfy the conditions (16.3) and (16.4) and so again solve the optimisation problem (16.2).

The above analysis demonstrates the asymptotic optimality of any strategy based on the use of a reservation parameter k, provided only that, in the limiting regime, k increases in accordance with (16.7). In practice, in a large network (here for large C), only a small value of k is required in order to achieve optimal performance. We also observe that the performance of a reservation parameter strategy is indeed robust with respect to variations in κ_1, κ_2.

This analysis also extends easily to the case where there are more than two call types, and also, with a little more difficulty, to that where the capacity constraint is of the form $\sum_{r\in\mathfrak{R}} A_r n_r \leq C$ for general positive integers A_r (see Bean *et al.* [4]). Here a different reservation parameter may be used for each call type, and, in the Kelly limiting scheme, a complete prioritisation and optimal control are again achieved asymptotically by allowing the differences between the reservation parameters to increase slowly.

16.3.2 Multiple resource models

Consider now the general case of the canonical model in which there is a set of resources $\mathcal{J}$ and in which state $\mathbf{n}$ of the network is subject to the constraints (16.1). Suppose that it is again desirable to choose admission controls so as to maximise the linear function $\phi(\mathbf{P}) := \sum_{r=1}^{R} a_r \kappa_r P_r$ of the stationary acceptance probabilities P_r, for given constants a_r, $r \in \mathcal{R}$. As in Section 16.3.1, we may consider the linear programming problem

$$\text{maximise } \phi(\mathbf{P}), \quad \text{subject to } P_r \in [0,1] \text{ for } r \in \mathcal{R}, \quad \sum_{r=1}^{R} A_{jr} \kappa_r P_r \le C_j \text{ for } j \in \mathcal{J}, \tag{16.8}$$

which provides an upper bound on the achievable values of the objective function ϕ. It is easy to see that this value may be asymptotically achieved within the Kelly limiting regime by reserving capacity $A_{jr}\kappa_r P_r$ at each resource j solely for calls of each type r, where here $\mathbf{P}$ is the solution of the problem (16.8). However this strategy is neither optimal in networks of finite capacity, nor is it robust with respect to variations in the parameters κ_r. At the opposite end of the spectrum from this complete partitioning policy is that of complete sharing. The latter can lead to unfairness if there are asymmetric traffic patterns, with the potential for some call types to receive better service than others. In practice it is expected that good strategies will be based on the sharing of resources and the use of reservation parameters—as was shown to be optimal for single resource networks in Section 16.3.1. (One example of a strategy midway between complete sharing and complete partitioning is virtual partitioning [35] [45]).

In the case of communications networks it is natural to allow also *alternative routing*, as described in the Introduction. An upper bound for the achievable performance is given by supposing that *repacking* is possible, i.e. that calls in progress may be rerouted as necessary. In this case, our model for the network reduces to an instance of the canonical model (as defined in the Introduction) with appropriately redefined set $\mathcal{J}$, matrix $A = (A_{jr})$ and capacities C_j. The upper bound on $\phi(\mathbf{P})$ given by the linear programming problem (16.8) is then also an upper bound in the more usual case in which repacking is not allowed. In the latter case practical control strategies are again based on the use of appropriate reservation parameters, and there is some hope that performance close to the upper bound above may be achieved in networks with sufficiently large capacities or sufficient diversity of routing, even without repacking. In applications reservation parameters are generally used to prioritise different traffic streams. In networks with alternative routing they also prevent the occurrence of network instabilities, where, for fixed parameter values, the network may have two or more relatively stable operating regimes—one in which most calls are directly routed, and others in which many calls are alternatively routed, with a resulting severe degradation of performance (see Gibbens [16], Kelly [27]). By giving priority to directly routed traffic, the use of reservation parameters prevents the network from slipping into an inefficient operating state.

There have been numerous investigations of control strategies for communications networks that employ either fixed or, particularly, alternative routing. Such strategies are often studied in the context of fully connected networks. Two of the most commonly studied are *least busy alternative* (LBA) routing and *dynamic alternative routing* (DAR). LBA routing seeks to route calls directly if possible, and otherwise routes them via that path which minimises the maximum occupancy on any of its links. Directly routed calls are usually "protected" with some form of reservation parameter (Kelly [27], Marbukh [33]). Hunt and Laws [23] showed that, for fully connected networks which permit only two-link alternative routes, LBA routing is asymptotically optimal in the diverse routing limit (see Section 16.4.5). This policy is robust to changes in traffic patterns, but has the difficulty that it requires information on the current states of all possible alternative paths before an alternative routing decision is made.

A much simpler routing scheme is DAR (Gibbens *et al.* [14], Gibbens and Kelly [15]). In this scheme, for each pair of nodes, a record is maintained of the current preferred alternative route, and this is the one that is used if a call cannot be routed directly. If neither the direct route nor the current preferred alternative route are available, then the call is rejected, and a new preferred alternative route is chosen at random from those available. Directly routed traffic is again usually protected by a reservation parameter. This policy is easy to implement. It does not require information about the current state of the system to be held at any node, just a record of the current preferred alternative route to other nodes. It is also robust to changes in traffic patterns—alternative routes on which the load increases will be discarded and replaced by routes on which the load is lower. Neither LBA routing nor DAR require traffic rates to be known or estimated (except approximately, in order to set the appropriate level of the reservation parameters).

Acceptance probabilities for controlled loss networks are usually estimated using a generalised version of the reduced load or knapsack approximation of Section 16.2.4. As there, we make the approximation (16.6) for the stationary distribution π' of the resource occupancy process $\mathbf{m}(\cdot)$. Each of the marginal distributions π'_j is estimated as the stationary distribution of a Markov process on $\{0,\dots,C_j\}$ which approximates the behaviour of the resource j considered in isolation. Let p_{jr} be the probability under this distribution that a call of type r is accepted, subject to the controls of the model, with $p_{jr} = 1$ if $A_{jr} = 0$. In the case of the canonical model, in which no alternative resource usage is allowed, calls of each type r are assumed to arrive at resource j at a rate $\nu_r \prod_{k \neq j} p_{kr}$—this is the "reduced load" for calls of type r at this resource; further, calls of this type arriving at this resource are subject to the acceptance controls of the model and, if accepted, depart at rate μ_r as usual. The estimated stationary distribution π'_j then determines the acceptance probabilities p_{jr} at the resource j. Thus we are again led to a set of fixed point equations which determine—not always uniquely—the acceptance probabilities p_{jr} for all $r \in \mathcal{R}$ and $j \in \mathcal{J}$. Finally the stationary network acceptance probability P_r for calls of each type r is again given by $P_r = \prod_{j \in \mathcal{J}} p_{jr}$.

In the case of a communications network where the canonical model is extended by allowing the possibility of alternative routing, it is necessary to modify the above approximation. Suppose, for example, that a link (resource) j forms part of the second choice route for calls of type r. Then, in the one-dimensional process associated with link j, the arrival rate for calls of type r is taken to be the product of the arrival rate ν_r at the network, the probability that a call of this type is rejected on its first-choice route, and (as before) the probabilities that the call can be accepted at each of the remaining resources on the alternative route (see e.g. Gibbens and Kelly [15]).

The basis of the reduced load approximation is the approximate factorisation of the distribution π' above. In the case of controlled networks, this approximation fails to become exact under the Kelly limiting regime in which capacities and arrival rates increase in proportion. It may, however, be expected to hold under sufficiently diverse routing. It is known to be remarkably accurate in most applications.

16.4 Dynamical behaviour and stability

We now consider the dynamical behaviour of large networks. As well as such behaviour being of interest in its own right—for example in networks in which input rates change suddenly, *fixed points* of network dynamics correspond to equilibrium, or quasi-equilibrium, states of the network (see below). The identification of such points is often the key to understanding long-term behaviour, in particular to resolving stability questions and determining stationary distributions where (as is usual) the latter may not be directly calculated. However, we note that it is characteristic of loss networks that, from any initial state, equilibrium is effectively achieved within a very few call holding times, so that transient *performance* is of less significance than is the case for networks which permit queueing.

16.4.1 Fluid limits for large capacity networks

We describe a theory first suggested by Kelly [27]. We yet again assume the Kelly limiting scheme described in the Introduction, in which the network topology is held fixed and arrival rates and capacities are allowed to increase in proportion. More explicitly, we consider a sequence of networks satisfying our usual Markov assumptions (though this is not strictly necessary) and indexed by a scale parameter N. All members of the sequence are identical in respect of the (finite) sets $\mathcal{R}$, $\mathcal{J}$, the matrix $A = (A_{jr}, j \in \mathcal{J}, r \in \mathcal{R})$, and the departure rates μ_r, $r \in \mathcal{R}$. For the Nth member of the sequence, calls of each type r arrive at rate $N\nu_r$ for some vector of parameters ν, and the capacity of each resource j is NC_j for some vector of parameters $\mathbf{C}$, where, for simplicity, we take each C_j to be integer-valued. As always, it is convenient to define $\kappa_r = \nu_r/\mu_r$ for each $r \in \mathcal{R}$.

We now describe the rules whereby calls are accepted. For each N, let $\mathbf{n}^N(t) = (n_r^N(t),\ r \in \mathcal{R})$, where $n_r^N(t)$ is the number of calls of type r in progress at time t. Define also the *free capacity* process $\bar{\mathbf{m}}^N(\cdot) = (\bar{m}_j^N(\cdot),\ j \in \mathcal{J})$ where each $\bar{m}_j^N(t) = NC_j - \sum_{r\in\mathcal{R}} A_{jr} n_r^N(t)$ is the free capacity of resource j at time t. A call of type r arriving at time t is accepted if and only if the free capacity $\bar{\mathbf{m}}^N(t-)$ of the system, immediately prior to its arrival, belongs to some acceptance region $\bar{\mathcal{A}}_r \subset \mathbb{Z}_+^J$. We take the acceptance regions $\bar{\mathcal{A}}_r$, $r \in \mathcal{R}$, to be independent of N, although, in a refinement of the theory, some dependence may be allowed. Note that, in a change from our earlier conventions, the acceptance regions $\bar{\mathcal{A}}_r$ are defined in terms of the *free* capacity of each system.

While the above description defines instances of the canonical model of the Introduction, more sophisticated controls, such as those involving the use of alternative routing in communications networks, may be modelled by the suitable redefinition of input streams and acceptance sets (see Hunt and Kurtz [22]).

For each N, define the normalised process $\mathbf{x}^N(\cdot) = \mathbf{n}^N(\cdot)/N$, which takes values in the space

$$X = \{\mathbf{x} \in \mathbb{R}_+^R \colon \sum_{r\in\mathcal{R}} A_{jr} x_r \le C_j \text{ for all } j \in \mathcal{J}\}. \tag{16.1}$$

Assume that, as $N \to \infty$, the initial state $\mathbf{x}^N(0)$ converges in distribution to some $\mathbf{x}(0) \in X$, which, for simplicity, we take to be deterministic. Then we might expect that the process $\mathbf{x}^N(\cdot)$ should similarly converge in distribution to a *fluid limit* process $\mathbf{x}(\cdot)$ taking values in the space X, with dynamics given by

$$x_r(t) = x_r(0) + \int_0^t (\nu_r \tilde{P}_r(u) - \mu_r x_r(u))du, \qquad r \in \mathcal{R}, \tag{16.2}$$

where, for each t, $\tilde{P}_r(t)$ corresponds to the limiting rate at which calls of each type r are being accepted at time t.

A rigorous convergence result is given by Hunt and Kurtz [22]. A somewhat technical condition (always likely to be satisfied in applications) is required on the acceptance sets $\bar{\mathcal{A}}_r$. However, the main difficulty is that in some, usually rather pathological, cases the limiting acceptance rates $\tilde{P}_r(t)$ may fail to be unique.

In many cases, though, it is possible to show that, for each r, there does exist a unique function P_r on X such that, for each t, we have $\tilde{P}_r(t) = P_r(\mathbf{x}(t))$. In general, the trajectories of the limit process $\mathbf{x}(\cdot)$ are then deterministic functions of their initial positions $\mathbf{x}(0)$. The fixed points $\hat{\mathbf{x}}$ of the limit process $\mathbf{x}(\cdot)$ are given by the solutions of

$$\nu_r P_r(\hat{x}) = \mu_r \hat{x}_r, \qquad r \in \mathcal{R}. \tag{16.3}$$

In the case of a single fixed point $\hat{\mathbf{x}}$, to which all trajectories of $\mathbf{x}(\cdot)$ converge, it may be shown that the stationary distribution of the original normalised process $\mathbf{x}^N(\cdot)$ converges to that concentrated on the single point $\hat{\mathbf{x}}$ (see Bean *et al.* [5]). Then in particular, for each r, $P_r(\hat{\mathbf{x}})$ is the limiting stationary acceptance probability for calls of type r. In the case of multiple fixed points, those which are locally stable correspond to “quasi-stationary” distributions of the process $\mathbf{x}^N(\cdot)$, i.e. regimes which are maintained over periods of time which are lengthy but finite.

16.4.2 Single resource networks

As the simplest non-trivial application of the above theory, we consider the case $J = 1$ of a single resource, for which equilibrium behaviour was described in Section 16.3.1. It is again convenient to write A_r for A_{1r} for each r, and similarly C for C_1. The technical condition referred to above on the acceptance sets $\bar{\mathcal{A}}_r \subseteq \mathbb{Z}_+$, here reduces to the requirement that, for each r, *either* $\bar m \in \bar{\mathcal{A}}_r$ for all sufficiently large $\bar m \in \mathbb{Z}_+$ (we let $\mathcal{R}^*$ denote the set of such r) *or* $\bar m \notin \bar{\mathcal{A}}_r$ for all sufficiently large $\bar m \in \mathbb{Z}_+$.

Here the functions P_r defined above always exist (see Hunt and Kurtz [22]). To identify them, define, for each $\mathbf{x} \in X$, the Markov process $\bar m_{\mathbf{x}}(\cdot)$ on $\mathbb{Z}_+$ with transition rates given by

$$\bar m \longrightarrow \begin{cases} \bar m - A_r & \text{at rate } \nu_r I_{\{\bar m \in \bar{\mathcal{A}}_r\}} \\ \bar m + A_r & \text{at rate } \mu_r x_r, \end{cases} \tag{16.4}$$

Let $\pi_{\mathbf{x}}$ be the stationary distribution of this process where it exists. Define $\bar X \subseteq X$ by

$$\bar X = \{\mathbf{x} \in X\colon \sum_{r\in\mathcal{R}} A_r x_r = C \text{ and } \pi_{\mathbf{x}} \text{ exists}\}. \tag{16.5}$$

(The set $\bar X$ may be thought of as consisting of those points in X for which the limiting dynamics are “blocking”.) Then, for $\mathbf{x} \in \bar X$, we have $P_r(\mathbf{x}) = \pi_{\mathbf{x}}(\bar{\mathcal{A}}_r)$ for all r; for $\mathbf{x} \in X \setminus \bar X$, we have $P_r(\mathbf{x}) = 1$ for $r \in \mathcal{R}^*$ and $P_r(\mathbf{x}) = 0$ for $r \notin \mathcal{R}^*$. The fixed points $\hat{\mathbf{x}}$ of the limiting dynamics (in general there may be more than one such) are then given by the solutions of (16.3).

Consider now the case of reservation-type controls, and suppose that the call types are arranged in order of decreasing priority. The acceptance regions are thus given by $\bar{\mathcal{A}}_r = \{\bar m\colon \bar m \geq k_r + A_r\}$ for some $0 = k_1 \leq k_2 \leq \cdots \leq k_R$ and we have $\mathcal{R}^* = \mathcal{R}$. It is easy to see that, in the *light traffic* case given by $\sum_{r\in\mathcal{R}} A_r \kappa_r \leq C$, the single fixed point $\hat{\mathbf{x}}$ of the limiting dynamics is given by $\hat x_r = \kappa_r$ for all r, and that all trajectories of these dynamics converge to $\hat{\mathbf{x}}$. In the *heavy traffic* case given by $\sum_{r\in\mathcal{R}} A_r \kappa_r > C$, define $\hat X \subseteq X$ by

$$\hat X = \{\mathbf{x} \in X\colon \sum_{r\in\mathcal{R}} A_r x_r = C \text{ and } x_r < \kappa_r \text{ for all } r \in \mathcal{R}\}.$$

Then it is straightforward to show that $\hat X \subseteq \bar X$ and that all fixed points of the limiting dynamics lie within $\hat X$ (see Bean *et al.* [4]). In the case where $A_r = 1$ for all r, it is also straightforward to show that there is a unique fixed point. It is unclear whether it is possible, for more general A_r, to have more than one fixed point.

Now define $r_0 \geq 0$ to be the maximum value of $r \in \mathcal{R}$ such that $\sum_{r \leq r_0} A_r \kappa_r \leq C$. Suppose that the reservation parameters $k_1, \dots, k_r$ are allowed to increase. Further consideration of the processes $\pi_{\mathbf{x}}$ shows that, in the limit (formally as these reservation parameters tend to infinity), the fixed point $\hat{\mathbf{x}}$ is necessarily unique and is such that $P_r(\hat{\mathbf{x}}) = 1$ for all $r \leq r_0$, with, in the heavy traffic case, $0 \leq P_{r_0+1}(\hat{\mathbf{x}}) \leq 1$ and $P_r(\hat{\mathbf{x}}) = 0$ for all $r \geq r_0 + 2$. Since the stationary distributions associated with our sequence of networks converge to that concentrated on the unique fixed point $\hat{\mathbf{x}}$, it follows that the reservation strategy does indeed approximate, and in the limit achieve, the complete prioritisation of call types discussed in Section 16.3.1. As mentioned there, and as easily verified from the above analysis, quite small values of the reservation parameters $k_1, \dots, k_r$ are sufficient to achieve a very good approximation to this prioritisation.

Even in the present single-resource case it is possible to achieve nonuniqueness of the fixed points of the limit process $\mathbf{x}(\cdot)$ by the use of more general, and sufficiently perverse, controls, in particular with the use of acceptance sets of the form $\bar{\mathcal{A}}_r = \{\bar{m} \colon A_r \leq \bar{m} \leq k_r + A_r\}$ for some $k_r \geq 0$ (see Bean *et al.* [5]). Thus we may construct networks which have several (very different) regimes which are quasi-stationary in the sense discussed above.

16.4.3 Multi-resource networks: the uncontrolled case

We now consider multi-resource networks, and again study the behaviour of the fluid limit process $\mathbf{x}(\cdot)$ associated with the Kelly limiting scheme. Here in general a rich variety of behaviour is possible. However, in the case of the uncontrolled networks of Section 16.2, in which calls of all types are accepted subject only to the availability of sufficient capacity, the process $\mathbf{x}(\cdot)$ is rather well-behaved. Note that here, in terms of the available free capacity, the acceptance sets are given by, for each $r \in \mathcal{R}$,

$$\bar{\mathcal{A}}_r = \{\bar{\mathbf{m}} \colon \ \bar{m}_j \geq A_{jr} \text{ for all } j\}. \tag{16.6}$$

Recall also that X is as given by (16.1). Define the (real-valued) concave function f on X by

$$f(\mathbf{x}) = \sum_{r \in \mathcal{R}} (x_r \log \nu_r - x_r \log \mu_r x_r + x_r) \tag{16.7}$$

and let $\hat{\mathbf{x}}$ be the value of $\mathbf{x}$ which maximises $f(\mathbf{x})$ in X. Kelly [26] shows that, as $N \to \infty$, the stationary distribution of the process $\mathbf{x}^N(\cdot)$ converges to that concentrated on the single point $\hat{\mathbf{x}}$. (Indeed this is the basis of his original derivation of the limiting acceptance probabilities considered in Section 16.2.4.)

Assume for the moment the unique existence of the functions P_r on X introduced above. Then, for the fluid limit process $\mathbf{x}(\cdot)$, it follows from (16.2) and (16.7) that $df(\mathbf{x}(t))/dt = g(\mathbf{x}(t))$ where the function g on X is given by

$$\begin{aligned} g(\mathbf{x}) &= \sum_{r\in\mathcal{R}} \frac{\partial f(\mathbf{x})}{\partial x_r}\Big(\nu_r P_r(\mathbf{x}) - \mu_r x_r\Big) \\ &= \sum_{r\in\mathcal{R}} \Big(\log \nu_r - \log \mu_r x_r\Big)\Big(\nu_r P_r(\mathbf{x}) - \mu_r x_r\Big). \end{aligned}$$

Analogously to the preceding section, for each $\mathbf{x} \in X$, the limiting acceptance probabilities $P_r(\mathbf{x})$ are given by consideration of the stationary distribution of a "free capacity" Markov process whose transition rates depend on $\mathbf{x}$. Some simple analysis of the equilibrium equations which define this stationary distribution (see Zachary [43]) now shows that $g(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in X$ with equality if and only if $\mathbf{x} = \hat{\mathbf{x}}$.

Thus the dynamics of the limit process $\mathbf{x}(\cdot)$ are such that, away from the point $\hat{\mathbf{x}}$, the function $f(\mathbf{x}(\cdot))$ is always strictly increasing. It thus acts as a Lyapunov function, ensuring that all trajectories of the process $\mathbf{x}(\cdot)$ converge to the single fixed point $\hat{\mathbf{x}}$. Indeed a rigorous application of the fluid limit theory of Hunt and Kurtz [22] (again see Zachary [43], for details) shows this result continues to hold even if the functions P_r on X are not uniquely defined (whether this can ever happen in the case of uncontrolled networks remains an open problem). The result therefore establishes an important stability property of uncontrolled networks, and guarantees that the stationary distribution describes the typical behaviour of the network.

16.4.4 Multi-resource networks: the general case

For general multi-resource networks, the fluid limit process $\mathbf{x}(\cdot)$ associated with the Kelly limiting scheme may fail to be unique, and may in particular exhibit multiple fixed points. We describe in some detail an elementary example, which is a simplification of one due to Hunt [21]. Suppose that $R = 3$, $J = 2$, and that the matrix A is given by

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}.$$

Thus in particular calls of types 1 and 2 each require capacity from a single resource, while calls of type 3 require capacity from both resources in the network. Suppose further that the (free capacity) acceptance sets are given by, for some $k_1, k_2 \geq 1$,

$$\bar{\mathcal{A}}_1 = \{\bar{\mathbf{m}}\colon 1 \leq \bar{m}_1 \leq k_1\}, \quad \bar{\mathcal{A}}_2 = \{\bar{\mathbf{m}}\colon 1 \leq \bar{m}_2 \leq k_2\}, \quad \bar{\mathcal{A}}_3 = \{\bar{\mathbf{m}}\colon \bar{m}_1 \geq 1, \bar{m}_2 \geq 1\}.$$

(As Hunt remarks, this is not entirely unrealistic: in more complex networks, operating under some form of alternative routing, certain resources may have calls of certain types routed over them precisely when the network is in general very busy.) Finally suppose that $\mu_r = 1$ for all r and that the vectors ν and $\mathbf{C}$ defined in Section 16.4.1 (each to be scaled by N for the Nth member of the sequence of networks) are given by $\nu = (\nu_1, \nu_2, \nu_3)$ and $\mathbf{C} = (C, C)$.

The process $\mathbf{x}(\cdot)$ takes values in the space $X = \{\mathbf{x} \in \mathbb{R}^3_+ : x_1 + x_3 \leq C,\ x_2 + x_3 \leq C\}$. Its dynamics may be determined through the fluid limit theory outlined above. For $\mathbf{x} \in X_0 := \{\mathbf{x} \in X \colon x_1 + x_3 < C,\ x_2 + x_3 < C\}$ (corresponding to limit points of the dynamics well away from the capacity constraints) the limiting acceptance probabilities are well-defined and given by

$$P_1(\mathbf{x}) = P_2(\mathbf{x}) = 0, \qquad P_3(\mathbf{x}) = 1. \tag{16.8}$$

For $\mathbf{x} \in X_1 := \{\mathbf{x} \in X \colon x_1 + x_3 = C,\ x_2 + x_3 < C\}$ and for $\mathbf{x} \in X_2 := \{\mathbf{x} \in X \colon x_1 + x_3 < C,\ x_2 + x_3 = C\}$ (corresponding in both cases to limit points of the dynamics such that only one capacity constraint is relevant) the limiting acceptance probabilities are again well-defined and given by consideration of a Markov process on $\mathbb{Z}_+$ as in the single resource case considered in Section 16.4.2. (For $\mathbf{x} \in X_1$, for example, it follows from the definition of $\bar{A}_2$ that the transition rates of this Markov process are as if $\nu_2 = 0$.) For $\mathbf{x} \in X_{12} := \{\mathbf{x} \in X \colon x_1 + x_3 = C,\ x_2 + x_3 = C\}$ it is necessary to consider also a "free capacity" Markov process on $\mathbb{Z}^2_+$.

In the case $\nu_3 \leq C$, these Markov processes all fail to possess stationary distributions and the limiting acceptance probabilities are given by (16.8) for all $\mathbf{x} \in X$. Thus the limit process $\mathbf{x}(\cdot)$ is as if $\nu_1 = \nu_2 = 0$ and all trajectories of this process are deterministic functions of their initial values and tend to the single fixed point $\hat{\mathbf{x}} = (0, 0, \nu_3)$.

The case $\nu_3 > C$ is more interesting. Here it is readily verified that the limit process $\mathbf{x}(\cdot)$ possesses no fixed points in X_0. Within X_1 consideration of the stationary distribution of the Markov process defined in Section 16.4.2 shows that there is a single fixed point $\mathbf{x}^{(1)} = (a_1, 0, C - a_1)$ for some a_1 which is independent of ν_2. Similarly within X_2 there is a single fixed point $\mathbf{x}^{(2)} = (0, a_2, C - a_2)$ for some a_2 which is independent of ν_1. However, within X_{12} the dynamics of the limit process $\mathbf{x}(\cdot)$ are not deterministic. It is further not difficult to show that all trajectories of $\mathbf{x}(\cdot)$ which avoid the set X_{12} tend deterministically to one of the two fixed points $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$ above (depending on whether the set X_1 or the set X_2 is hit first). Those trajectories of $\mathbf{x}(\cdot)$ which do hit X_{12} may, in an appropriate probabilistic sense, tend to either $\mathbf{x}^{(1)}$ or $\mathbf{x}^{(2)}$.

The interpretation of the above behaviour is the following. Suppose that N is large and that, for example, resource 1 fills to capacity first. Then this resource remains full and in general blocks sufficient of the type 3 calls to ensure that resource 2 remains only partially utilised, with few or no calls of type 2 being accepted. This corresponds to a "quasi-stationary" state whose limit, as $N \to \infty$, is concentrated on the fixed point $\mathbf{x}^{(1)}$. Alternatively, if resource 2 fills to capacity first, the network settles, for an extended period of time, to a quasi-stationary state whose limit is concentrated on the fixed point $\mathbf{x}^{(2)}$. While, for finite N, transitions between these two quasi-stationary states will eventually occur, the time taken to do so can be shown to increase exponentially in N.

We illustrate this with an example. Let $C = 500$, $\nu_1 = \nu_2 = 200$, $\nu_3 = 600$, $\mu_1 = \mu_2 = \mu_3 = 1$, with $k_1 = k_2 = 4$. The fixed points under the Kelly limiting regime are given by $\mathbf{x}^{(1)} = (125, 0, 375)$ and $\mathbf{x}^{(2)} = (0, 125, 375)$, provided k_1, k_2 are scaled appropriately. Figure 16.1 plots $\bar{m}_2$, the free capacity on link 2 against $\bar{m}_1$, the free capacity on link 1 for two simulated sample paths of this process, both initially started at $n_1 = n_2 = n_3 = 0$. The free capacity on both links decreases rapidly as initially only 2-link calls are accepted into the network. Once the threshold at which 1-link calls are accepted is reached, the sample paths then typically move rapidly towards one or other of the quasi-stationary modes. In Figure 16.1 the two sample paths illustrate both of these behaviours. Note that it is apparent that in this example the modes do not coincide with the fixed points of the limiting regime – here k_1 and k_2 are not sufficiently large to permit that. However, taking larger k_1 and k_2 (here $k_1 = k_2 = 10$ appears to be sufficient), will ensure that the modes of the two quasi-stationary distributions coincide approximately with the two fixed points $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$. Figure 16.2 is a similar plot for two sample paths, but for a rescaled version of the system, with $C = 1000$, $\nu_1 = \nu_2 = 400$, $\nu_3 = 1200$, $\mu_1 = \mu_2 = \mu_3 = 1$, and $k_1 = k_2 = 8$. The fixed points under the Kelly limiting regime are then given by $\mathbf{x}^{(1)} = (250, 0, 750)$ and $\mathbf{x}^{(2)} = (0, 250, 750)$, and we see that the quasi-stationary

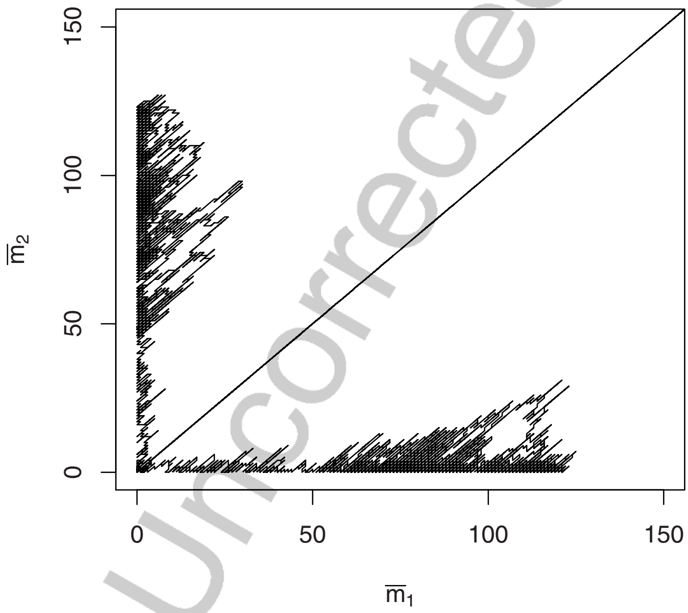


Fig. 16.1: $\bar{m}_2$ vs. $\bar{m}_1$ plotted for two simulated sample paths when $C = 500$, $\nu_1 = \nu_2 = 200$, $\nu_3 = 600$, $\mu_i = 1$, $i = 1,2,3$ and $k_1 = k_2 = 4$, $n_i(0) = 0$, $i = 1,2,3$. The simulations have been run for 10 time units.

distributions are now more nearly centred about these. With a further scaling to $C = 1500$ (not shown here) the Kelly limiting regime fixed points give a very good fit to the modes of the quasi-stationary distributions.

The behaviour in the above example is typical of that which may occur in more general networks—in particular those using alternative routing strategies—which are poorly controlled. Fluid limits may be used to study behaviour in networks with high capacities and correspondingly high arrival rates, and to choose values of, for example, reservation parameters so as to ensure that the network does not spend extended periods of time in states in which it is operating inefficiently. A realistic example here is the fully-connected network with alternative routing considered in Section 16.3.2. Gibbens and Kelly [15] and Gibbens *et al.* [16] give examples of the accuracy of this approximation, both with and without trunk reservation controls, primarily for overloaded networks (where DAR would actually be in use). In their examples the approximation performs worst when no controls are in place (Gibbens and Kelly [15] cite an example of a network with 10% overload, where

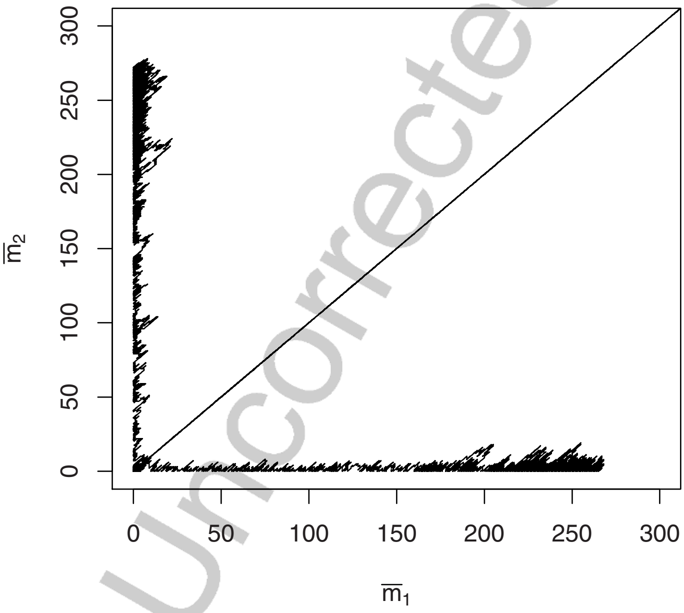


Fig. 16.2: $\bar{m}_2$ vs. $\bar{m}_1$ plotted for two simulated sample paths when $C = 1000$, $\nu_1 = \nu_2 = 400$, $\nu_3 = 1200$, $\mu_i = 1$, $i = 1,2,3$ and $k_1 = k_2 = 8$, $n_i(0) = 0$, $i = 1,2,3$. The simulations have been run for 10 time units.

the approximation has an error of less than 10%, based on simulation results), and correspondingly better as the trunk reservation control is increased.

As noted above, fluid limits may also be used to study equilibrium behaviour, especially in the case where all trajectories of the limit process $\mathbf{x}(\cdot)$ tend to a unique fixed point $\hat{\mathbf{x}}$. In particular we may show that, for the Kelly limiting regime considered here, the limiting stationary distribution of the free capacity processes $\bar{\mathbf{m}}^N(\cdot)$ in general only has a product form in the case of uncontrolled networks. This product-form assumption is the basis of the commonly used approximations considered in Section 16.3.2. Its justification owes more to the results for the diverse routing limit also considered there and in Section 16.4.5.

16.4.5 The diverse routing limit

In this section we consider the fluid limit obtained under the diverse routing regime discussed in the Introduction. Although a high degree of symmetry is required in order to obtain formal limits, the results obtained lend support to the commonly made assumptions of independence of resource blocking which are used, for example, in the construction of the approximations discussed in Section 3.2.

As outlined earlier, the diverse routing regime holds when the numbers of resources and possible "routes" in the network increase, while the total capacity and arrival rate at each resource remains constant. For this limit to exist we require a high degree of symmetry in the network. There are two canonical examples (with variants) that have been extensively studied. We describe both here using the terminology of communications networks.

The first is the so-called *star network* (see, for instance, Whitt [42], Ziedins and Kelly [47], Hunt [20]). Here there are K links, each with capacity C. The scale parameter of the regime is then taken to be K. Assume that calls of any *size* $r \geq 1$ require unit capacity at each of r resources and have holding times with unit mean. Then in a symmetric network there are $\binom{K}{r}$ possible choices of the set of links for such a call. Let the arrival rate for each such choice be $\nu_r^K = \nu_r / \binom{K-1}{r-1}$, so that the total arrival rate at each resource for calls of size r is exactly ν_r. For example, we may assume that the K links are distributed around a central hub, through which all communications must pass. Many variants of this model are possible—multiple call sizes can coexist in the network, as can multiple capacities, provided only that the proportion of links with any given capacity remains constant as K increases. The network is assumed to have fixed routing and the only permissible controls are those on admission.

Let $\mathbf{x}^K(t) = (x_j^K(t), j \in \mathcal{J})$ where $x_j^K(t)$ is the proportion of links in which j units of capacity are in use at time t. For the network without admission controls, Whitt [42] showed that, given the weak convergence of the initial points $\mathbf{x}^K(0)$ to $\mathbf{x}(0)$, the process $\mathbf{x}^K(\cdot)$ converges weakly to a deterministic limit process $\mathbf{x}(\cdot)$, which satisfies a set of first-order differential equations with a unique fixed point $\hat{\mathbf{x}}$, such that $\mathbf{x}(t) \to \hat{\mathbf{x}}$ as $t \to \infty$ for all initial $\mathbf{x}(0)$. The limit $\hat{\mathbf{x}}$ coincides exactly with that given by the Erlang fixed point approximation. Recall that the latter is obtained from the assumption that the stationary free capacity distributions on the various links of the network are independent of each other. For the case where all calls are of size two, Hunt [20] obtained a functional central limit theorem for the process $\mathbf{x}^K(\cdot)$, with the limit an Ornstein-Uhlenbeck diffusion process (as previously conjectured by Whitt), which was then extended to more general sizes and initial conditions by Graham and Meleard [18]. In the case of networks with admission controls very little has been proved. MacPhee and Ziedins [32] studied such networks and gave a weak convergence result for the process $\mathbf{x}^K(\cdot)$. However, there remain many open questions about the behaviour of this process.

The second canonical example of the diverse routing regime is that of the fully connected network with alternative routing (Hunt and Laws [23]). Here both admission and routing controls are possible. The network has N nodes; between each pair of these there is a link with capacity C, so that the total number of links is $K = \binom{N}{2}$. Here again K is the scale parameter. Calls arrive at each link at rate ν; each call has a unit capacity requirement and holding time of mean 1. There are three possible actions on the arrival of a call: (i) accept the call at that link, (ii) select a pair of links that form an alternative route between that pair of nodes and route the call along this, or (iii) reject the call. Hunt and Laws showed that an asymptotically optimal policy, in the sense of minimising the average number of lost calls in equilibrium, is to route a call directly if possible and otherwise to route it via an alternative route, provided that the remaining free capacity on each link of the alternative route is at least some reservation parameter k, where the optimal choice of k is determined by the parameters K and ν. The optimal choice of alternative route is given by choosing that which is least busy, i.e. which maximises the minimum of the free capacities on the two links. The analysis of Hunt and Laws largely dispenses with the graph structure inherent in the choice of alternative routes, an assumption justified by analogy with earlier results of Crametz and Hunt [11] in relation to the simpler model without reservation – see also Graham and Meleard [17], who show a propagation of chaos result for this system.

As in the example of the star network, of interest here is the process $\mathbf{x}^K(\cdot)$, defined as earlier. Hunt and Laws showed weak convergence of this process to a deterministic limit process. They showed that this limit process satisfies differential equations which yield the constraints for a linear programming problem, the solution to which gives an upper bound on the acceptance probabilities. (These constraints correspond to the detailed balance equations that in equilibrium govern the changes in occupancy of a single link.) They further showed that their policy achieves this upper bound.

16.5 Further developments and open questions

Our discussion has of necessity omitted many topics of interest, some of which we mention briefly here, as well as discussing some remaining open questions.

One such topic is the application of large deviations techniques to loss networks in order to estimate, for example, blocking probabilities in cases where it is important to keep these very small. For an excellent introduction to this see, for instance, Shwartz and Weiss [40]; later papers include those by Simonian *et al.* [41] and by Graham and O'Connell [19].

Another important topic is that of diffusion approximations, which appear briefly in Gibbens *et al.* [16], were mentioned in Section 16.4.5 and have been studied by others, including Puhalskii and Reiman [37] and Knessl and Morrison [29]. Much work has also been done on refinements of approximations adapted to particular situations, computational techniques for loss networks (see e.g. Louth *et al.* [30] and Choudhury *et al.* [9]), and bounds on blocking probabilities (Boucherie and van Dijk [6] is a recent example of the latter).

In some models of communications networks, particularly those whose graph structure is tree-like, the network topology may be such as to lend itself to more accurate calculations of acceptance probabilities, involving recursions that do not make the link independence assumption (16.6) that is such an essential feature of the approximations presented above (see Zachary and Ziedins [46]).

We have not directly addressed here the solution of the numerous optimization problems associated with loss networks, including network design and capacity requirements and the use of pricing mechanisms for control.

Extensions of loss network models include recent work by Antunes *et al.* [3] which studies a variant of the model where customers may obtain service sequentially at a number of resources, each of which is a loss system. The aim here is to model a cellular wireless system where a call in progress may move from base station to base station. Several authors have also considered explicitly systems with time-varying arrival rates and/or retries (see, for example, Jennings and Massey [24], Massey and Whitt [34] and Abdalla and Boucherie [1]).

A large number of interesting and important open problems remain. The approach to most of these seems to lie in a better understanding of network dynamics. There has been no systematic investigation of how to achieve asymptotically optimal control in a general network (for example in the sense of Section 16.3.2), using controls which are simple, decentralised, and robust with respect to variations in network parameters, although, for communications networks, there is a belief that this will usually combine some form of alternative routing with the use of reservation parameters to guarantee stability.

A further major problem is that of the identification of instability, where the state of a network may remain over extended periods of time in each of a number of "quasi-equilibrium" distributions, some of which may correspond to highly inefficient performance. Instability is further closely linked to problems of phase transition in the probabilistic models of statistical physics, and to the study of how phenomena such as congestion propagate through a network. At present such results as exist are mostly for very regular network topologies (see, for instance, Ramanan *et al.* [38] and Luen *et al.* [31])—but see also Antunes *et al.* [2, 3] for examples of such behaviour in heterogeneous communications networks.

Questions related to those above concern the identification of fluid limits, and in particular the problem of the uniqueness of their trajectories given initial conditions. It is notable that the uniqueness question has not yet been resolved even in the case of a general *uncontrolled* loss network, although it is known that here all trajectories do tend to the same fixed point, thus guaranteeing network stability. Further, while fixed points of fluid limits identify quasi-equilibrium states of a network, detailed behaviour within such states, and the estimation of the time taken to pass between them, requires a more delicate analysis based on the study of diffusion limits. Here relatively little work has been done (see Fricker *et al.* [13]).

Finally we mention that loss networks may be seen as a subclass of a more general class of stochastic models, with state space $\mathbb{Z}_+^R$ for some R and fairly regular transition rates between neighbouring states. Notably their analysis has much in common with that of processor-sharing networks in which calls again have a simultaneous resource requirement (see the chapter on processor-sharing networks in the present volume). A unified treatment is still awaited.

References

1. Abdalla, N., Boucherie, R.J.: Blocking probabilities in mobile communications networks with time-varying rates and redialing subscribers. Ann. Oper. Res. **112**, 15–34 (2002)
2. Antunes, N., Fricker, C., Robert, P., Tibi, D.: Metastability of CDMA cellular systems. In: Proc. MOBICOM 2006, pp. 206 –214. ACM, New York (2006)
3. Antunes, N., Fricker, C., Robert, P., Tibi, D.: Stochastic networks with multiple stable points. Ann. Probab. **36**, 255–278 (2008)
4. Bean, N.G., Gibbens, R.J., Zachary, S.: Asymptotic analysis of large single resource loss systems under heavy traffic, with applications to integrated networks. Adv. Appl. Probab. **27**, 273–292 (1995)
5. Bean, N.G., Gibbens, R.J., Zachary, S.: Dynamic and equilibrium behaviour of controlled loss networks. Ann. Appl. Probab. **7**, 873–885 (1997)
6. Boucherie, R.J., van Dijk, N.M.: Monotonicity and error bounds for networks of Erlang loss queues. Queueing Syst. **62**, 159–193 (2009)
7. Brockmeyer, E., Halstrom, H.L., Jensen, A.: The Life and Works of A.K. Erlang. Academy of Technical Sciences, Copenhagen (1948)
8. Burman, D.Y., Lehoczky, J.P., Lim, Y.: Insensitivity of blocking probabilities in a circuit switching network. J. Appl. Probab. **21**, 850–859 (1984)
9. Choudhury, G.L., Leung, K.K., Whitt, W.: An algorithm to compute blocking probabilities in multirate multiclass multi-resource loss models. Adv. Appl. Probab. **27**, 1104–1143 (1995)
10. Chung, S.P., Ross, K.W.: Reduced load approximations for multirate loss networks. IEEE T. Commun. **41**, 1222–1231 (1993)
11. Crametz, J.P., Hunt, P.J.: A limit result respecting graph structure for a fully connected loss network with alternative routing. Ann. Appl. Probab. **1**, 436–444 (1991)
12. Dziong, Z., Roberts, J.W.: Congestion probabilities in a circuit-switched integrated services network. Perform. Evaluation **7**, 267–284 (1987)
13. Fricker, C., Robert, P., Tibi, D.: A degenerate central limit theorem for single resource loss systems. Ann. Appl. Probab. **13**, 561-575 (2003)
14. Gibbens, R.J., Kelly, F.P., Key, P.B.: Dynamic alternative routing—modelling and behaviour. In: Bonatti, M.(ed.) Proc. 12th Int. Teletraffic Congress. Elsevier, Turin (1989)
15. Gibbens, R.J., Kelly, F.P.: Dynamic routing in fully connected networks. IMA J. Math. Control and Inf. **7**, 77–111 (1990)

16. Gibbens, R.J., Hunt, P.J., Kelly, F.P.: Bistability in communication networks. In: Grimmett, G., Welsh, D. (eds.) Disorder in Physical Systems: a Volume in Honour of John M. Hammersley, pp. 113–127. Oxford University Press (1990)
17. Graham, C. and Meleard, S.: Propagation of chaos for a fully connected loss network with alternate routing. Stoch. Proc. Appl. **44**, 159-180 (1993)
18. Graham, C., Meleard, S.: Dynamic asymptotic results for a generalized star-shaped loss network. Ann. Appl. Probab. **5**, 666–680 (1995)
19. Graham, C., O'Connell, N.: Large deviations at equilibrium for a large star-shaped loss network. Ann. Appl. Probab. **10**, 104–122 (2000)
20. Hunt, P.J.: Loss networks under diverse routing: the symmetric star network. Adv. Appl. Probab. **25**, 255–272 (1995)
21. Hunt, P.J.: Pathological behaviour in loss networks. J. Appl. Probab. **32**, 519–533 (1995)
22. Hunt, P.J., Kurtz, T.G.: Large loss networks. Stoch. Proc. Appl. **53**, 363–378 (1994)
23. Hunt, P.J., Laws, C.N.: Asymptotically optimal loss network control. Math. Oper. Res. **18**, 880–900 (1993)
24. Jennings, O.B., Massey, W.A.: A modified offered load approximation for nonstationary circuit switched networks. Telecommunication Systems **7**, 229–251 (1997)
25. Kaufman, J.S.: Blocking in a shared resource environment. IEEE T. Commun. **29**, 1474–1481 (1981)
26. Kelly, F.P.: Blocking probabilities in large circuit-switched networks. Adv. Appl. Probab. **18**, 473–505 (1986)
27. Kelly, F.P.: Loss networks. Ann. Appl. Probab. **1**, 319–378 (1991)
28. Kelly, F.P., Key, P., Zachary, S.: Distributed admission control. IEEE J. Sel. Areas Commun. **18**, 2617–2628 (2000)
29. Knessl, C., Morrison, J.A.: A two-dimensional diffusion approximation for a loss model with trunk reservation. SIAM J. Appl. Math. **69**, 1457–1476 (2008)
30. Louth, G., Mitzenmacher, M., Kelly, F.: Computational complexity of loss networks. Theor. Comput. Sci. **125**, 45–59. (1994)
31. Luen, B., Ramanan, K., Ziedins, I.: Nonmonotonicity of phase transitions in a loss network with controls. Ann. Appl. Probab. **16**, 1528–1562 (2006)
32. MacPhee, I.M., Ziedins, I.: Admission controls for loss networks with diverse routing. In: Kelly, F.P., Zachary, S., Ziedins, I. (eds.) Stochastic Networks: Theory and Applications, pp. 205–214. Oxford University Press (1996)
33. Marbukh, V.: Loss circuit switched communication network–performance analysis and dynamic routing. Queueing Syst. **13**, 111–141 (1993)
34. Massey, W.A., Whitt, W.: An analysis of the modified offered-load approximation for the nonstationary loss model. Ann. Appl. Probab. **4**, 1145-1160 (1994)
35. Mitra, D., Ziedins, I.: Virtual partitioning by dynamic priorities: a technique for fairly and efficiently sharing a resource between several services. In: Plattner, B. (ed) Broadband Communications, Lecture Notes in Computer Science 1044, pp. 173–185. Springer (1996)
36. Pechinkin, A.V.: A new proof of Erlang's formula for a lossy multichannel queueing system. Soviet J. Comput. System Sci. **25**, 165–168 (1987)
37. Puhalskii, A.A.,Reiman, M.I.: A critically loaded multirate link with trunk reservation. Queueing Syst. **28**, 157–190 (1998)
38. Ramanan, K., Sengupta, A., Ziedins, I., Mitra, P.: Markov random field models of multicasting in tree networks. Adv. Appl. Probab. **34**, 58–84 (2002)
39. Ross, K.W.: Multiservice loss models for broadband telecommunication networks. Springer, New York (1995)
40. Shwartz, A., Weiss, A.: Large deviations for performance analysis. Chapman and Hall, London (1994)
41. Simonian, A., Roberts, J.W., Theberge, F., Mazumdar, R.: Asymptotic estimates for blocking probabilities in a large multi-rate loss network. Adv. Appl. Probab. **29**, 806–829 (1997)
42. Whitt, W.: Blocking when service is required from several facilities simultaneously. AT&T Tech. J. **64**, 1807–185 (1985)

43. Zachary, S.: Dynamics of large uncontrolled loss networks. J. Appl. Probab. **37**, 685–695 (2000)
44. Zachary, S.: A note on insensitivity in stochastic networks. J. Appl. Probab., **44**, 238–248 (2007)
45. Zachary, S., Ziedins, I.: A refinement of the Hunt-Kurtz theory of large loss networks, with an application to virtual partitioning. Ann. Appl. Probab. **12**, 1–22 (2002)
46. Zachary, S., Ziedins, I.: Loss networks and Markov random fields. J. Appl. Probab. **36**, 403–414 (1999)
47. Ziedins, I.B., Kelly, F.P.: Limit theorems for loss networks with diverse routing. Adv. Appl. Probab. **21**, 804–830 (1989)

Chapter 17
A Queueing Analysis of Data Networks

Thomas Bonald and Alexandre Proutière

Abstract In packet-switched networks, resources are typically shared by a dynamic set of data flows. This dynamic resource sharing can be represented by a queueing network with state-dependent service rates. For a specific resource allocation we refer to as *balanced fairness*, the corresponding queueing network is a Whittle network and has an explicit stationary distribution. We give some key properties satisfied by balanced fairness and compare the resulting throughput performance to those obtained under the max-min fair and proportional fair allocations.

17.1 Introduction

Since Erlang's work on telephone networks at the beginning of the 20th century, queueing theory and communication networking have enjoyed a remarkable degree of cross-stimulation. Communication networks have been the source of interesting problems for queueing theorists; the developed theory has in turn proved very useful for the optimization of communication networks.

While the focus has long been on *circuit-switched* networks, the success of Ethernet technology and the subsequent rapid spread of the Internet have raised new issues specific to *packet-switched* networks. This is best illustrated by the decentralized nature of the associated control mechanisms. For instance, the throughput of each data flow is regulated by the congestion control algorithms of the transmission control protocol, TCP, implemented by the end hosts. Despite considerable research efforts, it remains unclear how these control mechanisms allocate the resources of such a large system as the Internet.

Thomas Bonald
Orange Labs, Issy-les-Moulineaux, France, e-mail: thomas.bonald@orange-ftgroup.com

Alexandre Proutière
Microsoft Research, Cambridge, UK, e-mail: alexandre.proutiere@microsoft.com

R.J. Boucherie and N.M. van Dijk (eds.), *Queueing Networks: A Fundamental Approach*,
International Series in Operations Research & Management Science 154,
DOI 10.1007/978-1-4419-6472-4_17, © Springer Science+Business Media, LLC 2011

The most fruitful approach in that respect was proposed by Kelly (see references in Section 17.13). It consists in representing each data flow as a fluid stream whose rate tends to the solution of a certain optimization problem, under the assumption that the set of active flows is constant. This assumption is reasonable provided the convergence rate of congestion control algorithms to equilibrium is much faster than the frequency of changes in the set of active flows. A question of interest concerns the fairness properties of the allocation at equilibrium. While *max-min fairness* has long been stated as an ideal objective, it turns out that current congestion control algorithms realize an allocation that is closer to *proportional fairness* where those flows that consume more resources tend to receive a lower bit rate.

The network dynamics that result from the evolution of the set of active flows can be studied for various allocations like max-min fairness and proportional fairness. These provide useful abstractions of the way packet-level control mechanisms allocate resources. A first issue is that of network stability: given the traffic intensity, does the number of active flows reach a finite steady state? The absence of admission control indeed leads a packet-switched network to congestion collapse in case of overload. In circuit-switched networks, on the other hand, the number of calls is naturally bounded. Another key issue concerns the throughput performance when the network is stable: what is the mean time required to transfer a document in steady state? Again, this question is irrelevant for circuit-switched networks where users experience quality of service through call blocking only. In packet-switched networks, the answer provides guidelines on how resources should be provisioned and allocated.

We shall see that queueing theory is instrumental in addressing these issues. Specifically, throughput performance can be evaluated for an allocation we refer to as *balanced fairness*, defined in such a way that the corresponding queueing network is a so-called Whittle network. Results provide a very good approximation of those obtained with proportional fairness. The considered fluid models of packet-switched networks under balanced fairness may in fact be considered as the analogue of standard models of circuit-switched networks such as the Erlang model. Both share the property that the stationary distribution of the network state is independent of all traffic characteristics beyond the traffic intensity. This *insensitivity* property is very useful in practice since it allows the development of simple engineering rules that do not require the knowledge of fine traffic statistics. It basically explains the enduring success of the Erlang formula, published in 1917 and still used to dimension today's telephone networks.

17.2 Capacity region

Consider a network that consists of L *resources*. Each resource may represent the capacity of a wireline link, the frequency band or transmission power of a wireless link, for instance. We denote by C_l the amount of resource l. A random set of data flows compete for access to these resources. Specifically, we consider an arbitrary number of N flow classes such that all flows within the same class require the same resources. There are x_i class-i flows and we refer to the corresponding line vector x as the *network state*. Each class-i flow requires an amount of resource l equal to A_{il} units per bit/s. If resource l represents the frequency band of a wireless link for instance, then each class-i flow requires A_{il} Hz per bit/s. The x_i class-i flows share evenly a total bit rate of ϕ_i bit/s. We denote by ϕ the corresponding line vector. The allocation must satisfy the component-wise inequality:

$$\phi A \leq C. \tag{17.1}$$

We refer to the set of vectors ϕ that satisfy this inequality as the *capacity region*. This is a convex polytope. In the following, we give a number of examples that illustrate the rich class of wireline and wireless networks covered by such linear capacity constraints.

Wireline networks

Consider a network that consists of L wireline links. The capacity of link l is C_l bit/s. Let $A_{il} = 1$ if class-i flows go through link l and $A_{il} = 0$ otherwise. Figures 17.1 and 17.2 show simple examples of such networks with their capacity region, respectively given by:

$$\begin{cases} \phi_1 + \phi_3 \leq 1, \\ \phi_2 + \phi_3 \leq 1, \end{cases} \quad \text{and} \quad \begin{cases} \phi_1 + \phi_2 + \phi_3 \leq 2, \\ \phi_1 \leq 1,\ \phi_2 \leq 1,\ \phi_3 \leq 1. \end{cases}$$

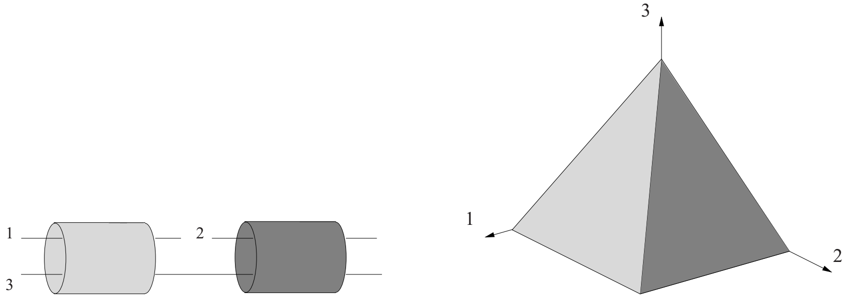


Fig. 17.1: A linear network and its capacity region.

Note that we do not specify the direction of the links. In Figure 17.1 for instance, the directions of both links may be either identical or opposite. In the former case, all classes represent usual point-to-point flows with a single source and a single destination. In the latter case, class 3 may represent point-to-multipoint flows with a single source, located between the two links, and two destinations.

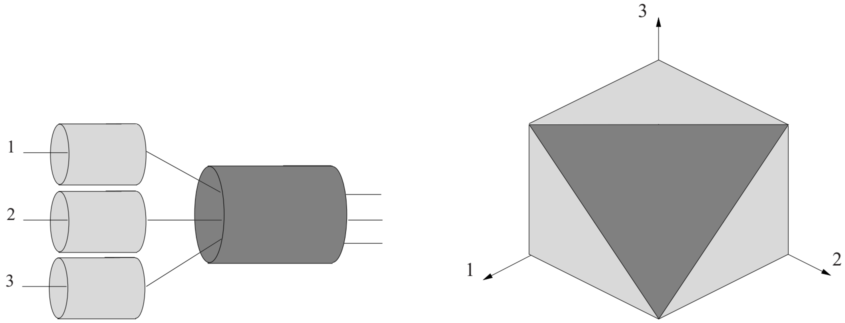


Fig. 17.2: A tree network and its capacity region.

Traffic splitting

The above model describes a network where each flow has a predetermined path in the network, possibly with several destinations. More complex routing schemes may be represented by linear capacity constraints. Assume for instance that the traffic generated by each class can be split over a predetermined set of paths. The capacity region is still a convex polytope, as illustrated by Figure 17.3 for $N = 3$ classes, where class-2 traffic is split over two paths. If the link used by class-3 flows has capacity 1 and the other two links have capacity 1/2, we get the capacity constraints:

$$\phi_1 \leq 1/2, \quad \phi_1 + \phi_2 \leq 1, \quad \phi_2 + \phi_3 \leq 1.$$

Note that the second capacity constraint may be viewed as a virtual link of capacity 2 used by class-1 and class-2 flows.

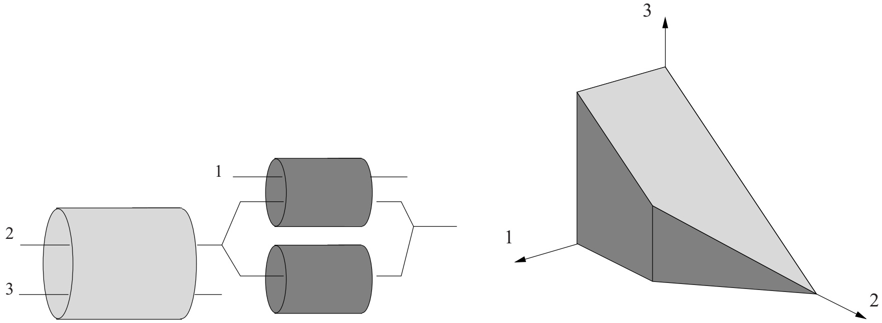


Fig. 17.3: A wireline network with traffic splitting and its capacity region.

Wireless networks

Modeling wireless networks is generally more difficult due to the joint frequency band and power allocation involved in the transmission. Consider the simple case of a wireless access point that transmits data to each active mobile one at a time, using the whole frequency band and the full power. Such a time-division multiplexing scheme is used for the downlink channel of standard third-generation cellular networks. Due to the short time-slot duration (typically less than 2 ms), the throughput of each mobile in fact depends mainly on the fraction of slots it receives, and not on the precise slot scheduling. In this setting, the capacity constraints of the system are also linear.

For instance, assume that a set of N modulation and coding schemes can be used by the mobiles depending on their radio conditions. We here assume that the radio conditions experienced by each mobile do not change during the data transfer so that each flow is transferred with a constant modulation and coding scheme. We refer to class-i flows as those flows that use the modulation and coding scheme i. Such flows have the bit rate c_i when served, so that ϕ_i/c_i is the fraction of time the access point serves a class-i flow. The capacity constraint is then given by:

$$\frac{\phi_1}{c_1} + \frac{\phi_2}{c_2} + \ldots + \frac{\phi_N}{c_N} \le 1,$$

as illustrated by Figure 17.4 for $N = 3$ modulation and coding schemes. The capacity region is a hyperplane.

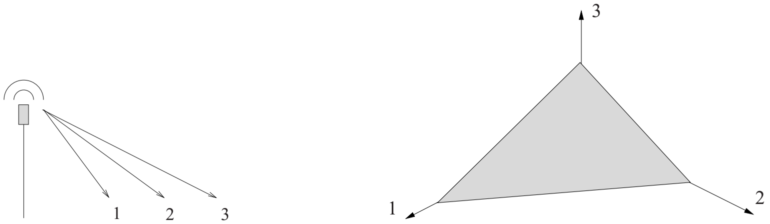


Fig. 17.4: A time-shared wireless access point and its capacity region.

Figure 17.5 shows the impact of the additional constraint of a wireline link of c bit/s, namely:

$$\phi_1 + \phi_2 + \ldots + \phi_N \le c.$$

More generally, the whole wireline backhaul network may be represented by accounting for the corresponding capacity constraints.

In the presence of several access points, the capacity region is typically non-convex due to interference. Consider the example of Figure 17.6 with three wireless access points. There are $N = 3$ flow classes, one per access point. For simplicity, we assume that all mobiles are co-located and thus experience the same radio con-

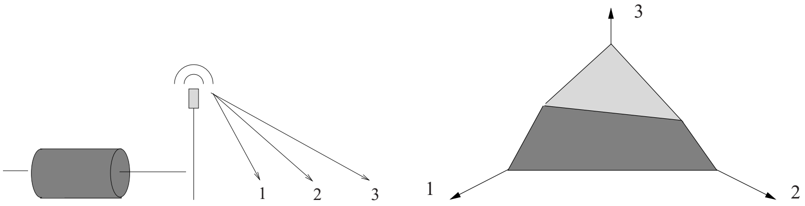


Fig. 17.5: A wireless access point with wireline backhaul and its capacity region.

ditions. We denote by P_i the power received by all mobiles from access point i, for $i = 1,2,3$. This received power cannot exceed a fixed value P. Let W be the available bandwidth and N_t be the thermal noise power. We use the Shannon formula as the bit rate function of the signal-to-interference-plus-noise ratio, which yields the following capacity region:

$$\phi_i \leq W \log_2 \left(1 + \frac{P_i}{N_t + \sum_{j \neq i} P_j} \right), \quad P_i \leq P, \quad i = 1,2,3.$$

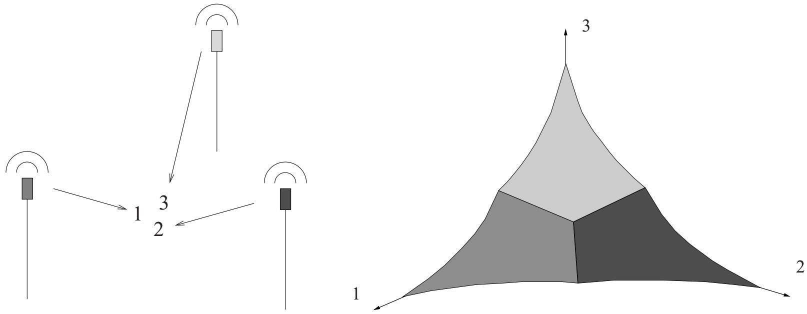


Fig. 17.6: A network of three interferring wireless access points and its capacity region.

As illustrated by Figure 17.6 for a signal-to-noise ratio $P/N_t = 1.5$, the capacity region is not convex. This is because the access points transmit simultaneously with a constant power. If the access points were transmitting at full power one at a time, which would require some form of coordination, interference would be cancelled and the capacity region would be the convex hull of that of Figure 17.6. Non-convex capacity regions raise specific issues, as discussed in Section 17.12.

Ad-hoc networks

Now consider a wireless network where mobiles cooperate in the sense that each mobile may relay the packets destined for another mobile. Figure 17.7 gives an example of such an ad-hoc network with 11 mobiles and $N = 3$ routes.

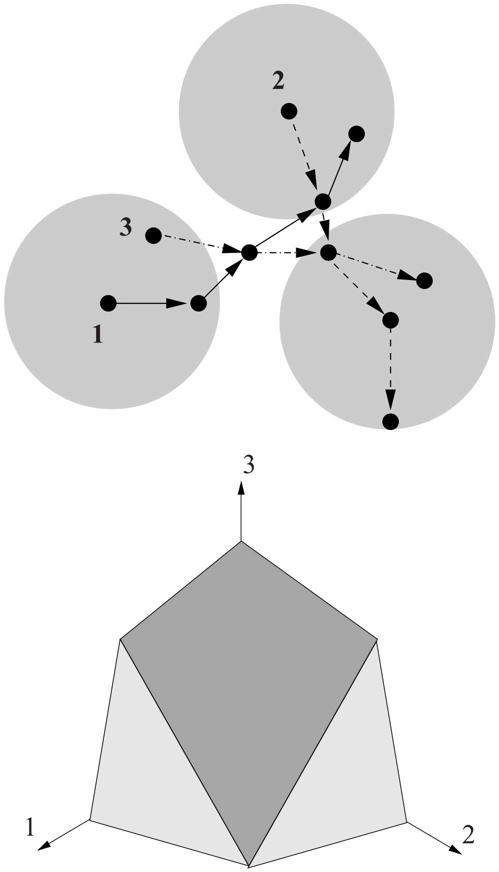


Fig. 17.7: An ad-hoc network and its capacity region.

We assume for simplicity that mobiles are static. Packets have a common fixed size and transmissions are synchronized. A mobile cannot send and receive at the same time. The transmission of a packet is successful if and only if the receiver lies in the transmission region of the sender and not in the transmission region of another transmitting mobile. We refer to any set of sender-receiver pairs that can be simultaneously active as a *transmission profile*, as illustrated by Figure 17.7 for three such pairs. Again, the transmission profiles are assumed to be scheduled at a sufficiently high frequency so that the allocation depends only on the fraction of time each transmission profile is used. The capacity region is then given by the convex hull of those rates obtained with a particular scheduling of the transmission profiles. For the network of Figure 17.7 for instance, the capacity constraints reduce to:

$$2\phi_1 + 2\phi_2 + 3\phi_3 \leq 1, \quad 3\phi_1 + \phi_2 + 2\phi_3 \leq 1, \quad \phi_1 + 3\phi_2 + 2\phi_3 \leq 1.$$

Note that, in practice, this capacity region is achieved by coordinating the transmission of the 11 mobiles. This may be realized by decentralized control algorithms provided some signaling information, not considered here, is exchanged by mobiles.

Flow rate limits

In addition to the global capacity constraints (17.1), flows may have individual rate constraints due for instance to the speed of the user's access line in wireline networks or the power constraint of the mobile in wireless networks. We denote by $a_i > 0$ the rate constraint of class-i flows in bit/s. We let $a_i = \infty$ if class-i flows do not have any individual rate constraint. Thus the total bit rate of class-i flows cannot exceed $x_i a_i$ in the presence of x_i class-i flows. Using vectorial notation, the allocation ϕ must satisfy the additional component-wise inequality:

$$\phi \leq xa, \tag{17.2}$$

where xa denotes the N-dimensional vector whose i-th component is equal to $x_i a_i$. Note that these rate constraints depend on the network state x and thus cannot be written in the form of some additional global capacity constraints (17.1). A single wireline link shared by $N = 2$ classes, with $a_1 < a_2$, is shown in Figure 17.8.

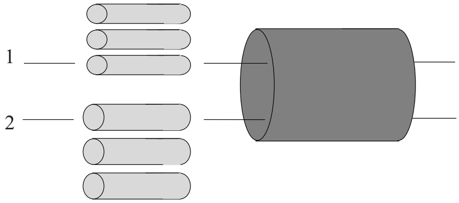


Fig. 17.8: A multirate system.

17.3 Traffic characteristics

We are interested in the steady-state behavior of the network state $x(t)$ that describes the number of ongoing flows of each class at time t. This depends both on the resource allocation and on traffic characteristics like the flow arrival process and the flow size distribution of each class. We assume that the vector ϕ of allocated bit rate depends on the network state x only and satisfies the capacity constraints (17.1) and (17.2) in all states. Max-min fairness, proportional fairness and balanced fairness correspond to specific allocation functions $\phi(x)$, described in Section 17.7.

Markovian setting

Consider the simple case where class-i flows arrive as a Poisson process of intensity λ_i and have i.i.d. sizes with exponential distribution of mean σ_i bits. The network state $x(t)$ then evolves as a Markov process. Specifically, let e_i be the unit vector with 1 in component i and 0 elsewhere. The transition rate from state x to state $x+e_i$ is equal to λ_i. Since the total bit rate of class-i flows is $\phi_i(x)$ in state x, the transition rate from state x to state $x-e_i$ is equal to $\phi_i(x)/\sigma_i$ for all states x such that $x_i > 0$. Provided $\phi_i(x) > 0$ for all states such that $x_i > 0$, which we assume in the following, the Markov process $x(t)$ is irreducible on $\mathbb{N}^N$.

Flow size distribution

The flow size distribution is typically not exponential but rather heavy-tailed in data networks. Informally stated, most flows consist of a few packets but most traffic is contained in a few large flows. We shall consider Cox distributions in the following, also known as phase-type distributions, that form a dense subset of the set of all distributions with non-negative support. This allows us to retain the Markovian description of the network state, but on a larger state space that includes the phases of the data transfers.

Session structure

Similarly, flows do not arrive as a Poisson process in practice. They are typically generated within sessions, each session being composed of a succession of flows separated by intervals of inactivity referred to as “think-times”. For instance, the second flow of a session starts once the first flow of the session is completed, after a think-time of random duration. The number of flows per session, the flow sizes and the think-time durations within each session have arbitrary distributions and may be correlated. Sessions, on the other hand, are mutually independent and are typically generated as a Poisson process. Again, we shall consider Cox distributions in the following so that the network state will still be described by a Markov process, but on a larger state space that includes all types of session and the phases of the corresponding flows and think-times.

17.4 Stability issues

A question of primary interest concerns the stability of the stochastic process $x(t)$ describing the evolution of the network state. We here do not make any specific assumption on traffic characteristics beyond stationarity and ergodicity.

Necessary condition

Let ρ_i be the traffic intensity of class-i flows in bit/s. This is the product of the arrival rate λ_i and the mean size σ_i of class-i flows. Clearly, a necessary condition for the network state $x(t)$ to reach a finite stationary regime is that the vector ρ of traffic intensities lies in the capacity region, that is if the following component-wise inequality is satisfied:

$$\rho A \leq C. \tag{17.1}$$

Property 17.4.1 *The above inequality is a necessary condition for stability.*

Proof. Assume that inequality (17.1) is violated. There exists a resource l such that:

$$\sum_i \rho_i A_{il} > C_l. \tag{17.2}$$

If l were the only resource, the system would correspond to a single server-queue with service capacity C_l, arrival rate $\lambda = \sum_i \lambda_i$ and mean service requirement:

$$\kappa = \sum_i \frac{\lambda_i}{\lambda} \sigma_i A_{il}.$$

The load $\lambda \kappa$ of this queue is larger than 1 in view of (17.2). The queue is unstable and, since the capacity constraints (17.1) include that of resource l, so is the original system.

Sufficient condition

It turns out that for usual allocations like max-min fairness, proportional fairness and balanced fairness, the necessary condition (17.1) is also sufficient, up to the critical case where the vector ρ lies on the boundary of the capacity region. Thus in the rest of the chapter, we assume that the following component-wise strict inequality is satisfied:

$$\rho A < C. \tag{17.3}$$

For max-min fairness and proportional fairness, the proof of stability is quite technical and requires some restrictive assumptions on the traffic characteristics. For balanced fairness, it is straightforward (cf. Proposition 17.7.1 below) and, in view of the insensitivity results derived in Section 17.8, valid for all traffic characteristics described in the previous section.

17.5 Flow throughput

We now assume that the network state $x(t)$ is stationary and ergodic and introduce a throughput measure, referred to as the *flow throughput*, that can be derived from its stationary distribution π. The flow throughput reflects the quality of data transfers as experienced by users in steady state in the following two senses.

Mean flow duration

The first definition is related to the mean flow duration. Specifically, the flow throughput is defined as the ratio of the mean flow size to the mean flow duration. We refer to the inverse of the flow throughput, namely the ratio of the mean flow duration to the mean flow size, as the *per-bit delay* (in s/bit). Let τ_i be the per-bit delay of class-i flows. Since the mean size of class-i flows is σ_i, the mean duration of class-i flows is equal to $\sigma_i\tau_i$ by definition. Denote by $\bar{x}_i$ the average number of class-i flows in steady state. By Little's law, we have:

$$\bar{x}_i = \lambda_i \times \sigma_i\tau_i = \rho_i\tau_i. \tag{17.1}$$

We deduce the flow throughput of class i:

$$\gamma_i = \frac{1}{\tau_i} = \frac{\rho_i}{\bar{x}_i}. \tag{17.2}$$

Mean instantaneous rate

The second definition corresponds to the mean instantaneous rate as experienced by users. Since the total bit rate allocated to class-i flows is evenly shared by these flows, the bit rate of a class-i flow is equal to $\phi_i(x)/x_i$ in any state x such that $x_i > 0$. Now the steady state probability that a class-i flow sees the network in state x is proportional to $x_i\pi(x)$ and therefore equal to $x_i\pi(x)/\bar{x}_i$. We deduce the flow throughput of class i:

$$\gamma_i = \sum_{x:x_i>0} \frac{x_i\pi(x)}{\bar{x}_i} \times \frac{\phi_i(x)}{x_i} = \frac{1}{\bar{x}_i} \sum_{x:x_i>0} \pi(x)\phi_i(x).$$

Expression (17.2) then follows from the traffic conservation equation:

$$\rho_i = \sum_{x:x_i>0} \pi(x)\phi_i(x),$$

which is a consequence of the ergodicity of the process $x(t)$.

Other throughput metrics

Clearly, a number of other performance metrics could be used to assess the quality of the data transfers. In the presence of per-flow rate constraints for instance, a quantity of interest is the probability that the instantaneous bit rate of some flow is less than its rate limit. A recursive algorithm to evaluate this probability is given in Section 17.9 in the case of a single wireline link.

17.6 Queueing analysis

Evaluating the flow throughput requires the derivation of the stationary distribution π. We first consider the Markovian setting described in Section 17.3 where flows of each class arrive as a Poisson process and have i.i.d. sizes with exponential distribution. The sensitivity of the results to these traffic assumptions depends on the allocation and will be discussed in Sections 17.7 and 17.8.

A queueing network

In the considered Markovian setting, the system may be viewed as a network of N parallel queues with state-dependent service rates. Specifically, class-i flows may be represented as customers that arrive at queue i as a Poisson process of intensity λ_i, have i.i.d. service requirements with exponential distribution of mean σ_i and leave the network once served. The traffic intensity at queue i is $\rho_i = \lambda_i \sigma_i$ (in bit/s). The number of customers present at queue i is x_i. The service rate ϕ_i of queue i, which corresponds to the bit rate allocated to class-i flows, depends on the network state x. This is illustrated in Figure 17.9 for $N = 2$ classes. Note that the vector of service rates ϕ is constrained by the capacity region (17.1) in all states x. Since the total bit rate allocated to each class is assumed to be evenly shared by the flows of this class, the service discipline of each queue is processor sharing.

Balance property

It turns out that the analysis of such a queueing network is intractable unless the service rates satisfy the following balance property:

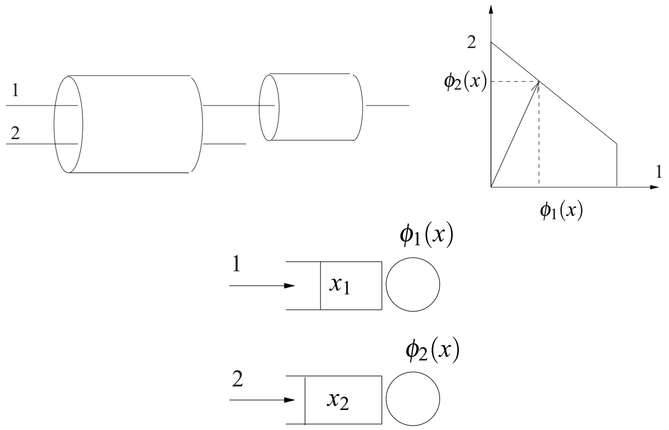


Fig. 17.9: A wireline network, its capacity region and the corresponding queueing network.

$$\forall i,j, \quad \forall x, \quad \phi_i(x)\phi_j(x-e_i) = \phi_j(x)\phi_i(x-e_j), \tag{17.1}$$

where we use the convention that $\phi(x) = 0$ if $x \notin \mathbb{N}^N$. Note that the balance property is equivalent to the reversibility of the Markov process $x(t)$, whose transition rates are given in Section 17.3. We refer to the corresponding queueing network as a Whittle network (cf. the appendix).

Stationary distribution

In view of the balance property (17.1), one can define a positive function Φ by $\Phi(0) = 1$ and:

$$\forall x \neq 0, \quad \Phi(x) = \frac{1}{\phi_{i_1}(x)\phi_{i_2}(x-e_{i_1})\dots\phi_{i_n}(e_{i_n})}, \tag{17.2}$$

where $x, x-e_{i_1}, x-e_{i_1}-e_{i_2}, \dots, e_{i_n}, 0$ denotes any direct path from state x to state 0. Conversely, the existence of such a function implies the balance property (17.1). Now let π be the positive measure on $\mathbb{N}^N$ defined (up to a multiplicative constant) by:

$$\forall x, \quad \pi(x) = \pi(0)\Phi(x)\rho^x, \tag{17.3}$$

where we use the notation:

$$\rho^x \equiv \prod_i \rho_i^{x_i}.$$

The measure π satisfies the detailed balance equations associated with the Markov process $x(t)$:

$$\forall x, \quad \pi(x)\phi_i(x)\sigma_i^{-1} = \pi(x-e_i)\lambda_i,$$

where we use the convention that $\phi_i(x) = 0$ if $x_i = 0$ and $\pi(x) = 0$ if $x \notin \mathbb{N}^N$. Thus $x(t)$ is indeed reversible, of invariant measure π. It is ergodic under the stability condition:

$$\sum_x \Phi(x)\rho^x < \infty, \tag{17.4}$$

in which case π is, after normalization, the stationary distribution of the network state.

17.7 Resource allocation

The balance property (17.1) is key to evaluating the stationary distribution of the network state and thus the flow throughput: if the resource allocation satisfies that property, there is a closed-form expression for the stationary distribution, which is insensitive to all traffic characterics beyond the traffic intensity (this is shown in the next section using the insensitivity property of Whittle networks); if the resource allocation does not satisfy the balance property, there is no closed-form expression for the stationary distribution, which is sensitive to all traffic characteristics (the corresponding queueing network is not a Whittle network).

Max-min fairness

The principle of max-min fairness is to allocate network resources as equally as possible without wasting resources. Max-min fairness is uniquely defined by the following *water-filling* procedure:

1. start from a bit rate equal to zero for all flows;
2. increase the bit rate of all flows at the same speed until the bit rate of some flows is constrained by the capacity region or by their rate limit; freeze the bit rate of these flows;
3. apply step 2 repeatedly to non-frozen flows until the bit rate of all flows is constrained by the capacity region or by their rate limit.

For the linear network of Figure 17.1 for instance, with equal link capacities and the same number of flows on each route, all flows have the same bit rate.

Max-min fairness does not satisfy the balance property (17.1) except if the capacity constraints (17.1) reduce to a single resource l and if all flows have the same resource requirement in the sense that $A_{il} = A_{jl}$ for all i, j. In the presence of additional per-flow constraints (17.2), all flows must also have the same rate limit.

Proportional fairness

Proportional fairness is based on the notion of *utility*. Specifically, assume the utility of a user for any data transfer is equal to the logarithm of his/her instantaneous bit rate. Proportional fairness is then defined as the unique allocation that maximizes the overall utility under the capacity constraints:

$$\forall x \neq 0, \quad \phi(x) = \arg \max_{\varphi:\, \varphi A \leq C, \varphi \leq xa} \sum_{i: x_i > 0} x_i \log\left(\frac{\varphi_i}{x_i}\right).$$

Coming back to the linear network of Figure 17.1 with equal link capacities and the same number of flows on each route, the bit rate of class-3 flows is half that of class-1 flows and class-2 flows: these flows that consume more resources receive a lower bit rate.

Proportional fairness does not satisfy the balance property in general. For linear networks like that of Figure 17.1, proportional fairness is balanced if and only if all links have the same capacity. For tree networks like that of Figure 17.2, proportional fairness coincides with max-min fairness and is not balanced. Like max-min fairness, proportional fairness is not balanced in the presence of additional per-flow constraints (17.2), except if the network reduces to a single resource and all flows have the same resource requirement and the same rate limit.

Balanced fairness

There is a unique allocation that satisfies the balance property and lies on the boundary of the capacity region. This allocation, that coincides with max-min fairness and proportional fairness in the particular cases where these allocations satisfy the balance property, is referred to as *balanced fairness*.

In view of (17.2), balanced fairness is uniquely defined by the corresponding balance function Φ as follows:

$$\forall x \neq 0, \quad \phi_i(x) = \frac{\Phi(x - e_i)}{\Phi(x)},$$

with the convention that $\Phi(x) = 0$ if $x \notin \mathbb{N}^N$. In view of the capacity constraints (17.1), the balance function must satisfy the inequalities:

$$\forall l, \quad \Phi(x) \geq \frac{1}{C_l} \sum_i A_{il} \Phi(x - e_i).$$

If the vector $\phi(x)$ lies on the boundary of the capacity set in all states $x \neq 0$, one of these inequalities must be an equality. We deduce that the balance function Φ is recursively defined by $\Phi(0) = 1$ and:

$$\forall x \neq 0, \quad \Phi(x) = \max_l \frac{1}{C_l} \sum_i A_{il} \Phi(x - e_i). \tag{17.1}$$

In the presence of per-flow rate limits (17.2), the balance function must satisfy the additional inequalities:

$$\forall i : x_i > 0, \quad \Phi(x) \geq \frac{1}{x_i a_i} \Phi(x - e_i).$$

The recursion becomes in this case:

$$\Phi(x) = \max \left\{ \max_l \frac{1}{C_l} \sum_i A_{il} \Phi(x - e_i), \max_{i:x_i>0} \frac{1}{x_i a_i} \Phi(x - e_i) \right\}. \tag{17.2}$$

In both cases, the stationary distribution of the network state is given by (17.3) under the stability condition (17.4). As mentionned in Section 17.4, this stability condition is in fact satisfied under the usual traffic conditions (17.3):

Property 17.7.1 *For balanced fairness, the network is stable if $\rho A < C$.*

Proof. Using (17.2), it may be easily verified by induction on $|x| \equiv \sum_i x_i$ that Φ is the smallest balance function that satisfies the capacity constraints in the sense that $\Phi(x) \leq \tilde{\Phi}(x)$ for all states x for any function $\tilde{\Phi}$ such that $\tilde{\Phi}(0) = 1$ and for all $x \neq 0$:

$$\forall l, \quad \sum_i A_{il} \frac{\tilde{\Phi}(x - e_i)}{\tilde{\Phi}(x)} \leq C_l, \quad \forall i : x_i > 0, \quad \frac{\tilde{\Phi}(x - e_i)}{\tilde{\Phi}(x)} \leq x_i a_i.$$

If $\rho A < C$, there is some vector $\tilde{\rho}$ which is component-wise strictly larger than ρ such that $\tilde{\rho} A < C$. In the absence of per-flow rate constraints, let $\tilde{\Phi}$ be the positive function defined by:

$$\tilde{\Phi}(x) = \prod_i \frac{1}{\tilde{\rho}_i^{x_i}}.$$

We have $\tilde{\Phi}(0) = 1$ and it follows from the inequality $\tilde{\rho} A < C$ that the capacity constraints are satisfied. We deduce that $\Phi(x) \leq \tilde{\Phi}(x)$ for all states x. In particular,

$$\sum_x \Phi(x) \rho^x \leq \sum_x \tilde{\Phi}(x) \rho^x = \sum_x \prod_i \left(\frac{\rho_i}{\tilde{\rho}_i} \right)^{x_i} < \infty.$$

The stability condition (17.4) is satisfied.

In the presence of per-flow rate constraints, the proof is similar with the balance function $\tilde{\Phi}$ defined as:

$$\tilde{\Phi}(x) = \prod_i \varphi_i(x_i),$$

where for each class i and all positive integers n,

$$\varphi_i(n) = \frac{1}{n!a_i^n} \quad \text{if } na_i \le \tilde{\rho}_i, \quad \varphi_i(n) = \frac{\varphi_i(n-1)}{\tilde{\rho}_i} \quad \text{otherwise.}$$

We have $\tilde{\Phi}(0) = 1$ and it may be easily verified that the capacity constraints are satisfied. The proof then follows from the fact that:

$$\sum_x \Phi(x)\rho^x \le \sum_x \tilde{\Phi}(x)\rho^x = \sum_x \prod_i \varphi_i(x_i)\rho_i^{x_i} < \infty.$$

17.8 Insensitivity results

In this and the following two sections, we focus on balanced fairness, for which analytical results can be derived. The throughput performance of max-min fairness and proportional fairness is compared to that of balanced fairness for various networks in Section 17.11. We here show that the stationary distribution of the network state under balanced fairness is independent of all traffic characteristics described in Section 17.3 beyond the traffic intensity.

Flow size distribution

We first assume that flows of each class arrive as a Poisson process and show that the stationary distribution (17.3) is insensitive to the flow size distribution[1]. We prove in addition that the flow throughput is independent of the flow size.

Consider the simple case of a Cox distribution that consists of a mixture of two exponential distributions. Specifically, class-i flows start with an exponential phase of mean $\sigma_{i,1}$ bits, which is followed by an exponential phase of mean $\sigma_{i,2}$ bits with probability p_i. The mean size of class-i flows is $\sigma_i = \sigma_{i,1} + p_i\sigma_{i,2}$. We denote by $\rho_{i,1} = \lambda_i\sigma_{i,1}$ the traffic intensity corresponding to the first phase, by $\rho_{i,2} = \lambda_i p_i \sigma_{i,2}$ the traffic intensity corresponding to the second phase. The total traffic intensity of class i is $\rho_i = \rho_{i,1} + \rho_{i,2}$.

Let $y = (y_1, y_2)$ where y_1 and y_2 are the vectors whose i-th component $y_{i,1}$ and $y_{i,2}$ gives the number of class-i flows in phases 1 and 2, respectively. Since the total bit rate allocated to class-i flows is evenly shared by these flows, the bit rate allocated to class-i flows in phases 1 and 2 is respectively given by:

$$\phi_{i,1}(y) = \phi_i(y_1 + y_2)\frac{y_{i,1}}{y_{i,1} + y_{i,2}} \quad \text{and} \quad \phi_{i,2}(y) = \phi_i(y_1 + y_2)\frac{y_{i,2}}{y_{i,1} + y_{i,2}}$$

in all states y such that $y_{i,1} + y_{i,2} > 0$. The corresponding balance property (17.1) is satisfied, so that the associated queueing network is a Whittle network (refer to the appendix). The balance function is:

[1] Recall that we restrict the analysis to Cox distributions. We refer the reader to [25] for the extension of this result to any distribution with finite mean.

$$y \mapsto \Phi(y_1+y_2)\prod_i \binom{y_{i,1}+y_{i,2}}{y_{i,1}}.$$

We deduce the stationary distribution of the number of flows of each class:

$$\begin{aligned}\forall x \neq 0, \quad \pi(x) &= \pi(0)\sum_{y:y_1+y_2=x} \Phi(y_1+y_2)\prod_i \binom{y_{i,1}+y_{i,2}}{y_{i,1}} \rho_{i,1}^{y_{i,1}}\rho_{i,2}^{y_{i,2}}\\ &= \pi(0)\Phi(x)\rho^x,\end{aligned}$$

which coincides with (17.3). Thus the stationary distribution of the number of flows of each class is insensitive to the chosen Cox distribution. We have the following additional insensitivity result.

Property 17.8.1 *For any class i, the flow throughput is the same for both phases of the flow and equal to γ_i.*

Proof. Let $\gamma_{i,1}$ and $\gamma_{i,2}$ be the flow throughput corresponding to phases 1 and 2 of class-*i* flows, respectively. Denoting by $\bar{y}_{i,1}$ and $\bar{y}_{i,2}$ the mean number of class-*i* flows in phases 1 and 2, respectively, we have in view of (17.2):

$$\gamma_{i,1} = \frac{\rho_{i,1}}{\bar{y}_{i,1}} \quad \text{and} \quad \gamma_{i,2} = \frac{\rho_{i,2}}{\bar{y}_{i,2}}.$$

The mean number of class-*i* flows in phase 1 is given by:

$$\begin{aligned}\bar{y}_{i,1} &= \pi(0)\sum_{y:y_{i,1}>0} y_{i,1}\Phi(y_1+y_2)\prod_j \binom{y_{j,1}+y_{j,2}}{y_{j,1}}\rho_{j,1}^{y_{j,1}}\rho_{j,2}^{y_{j,2}},\\ &= \pi(0)\sum_{y:y_{i,1}>0} (y_{i,1}+y_{i,2})\rho_{i,1}\Phi(y_1+y_2)\binom{y_{i,1}+y_{i,2}-1}{y_{i,1}-1}\rho_{i,1}^{y_{i,1}-1}\rho_{i,2}^{y_{i,2}}\\ &\qquad \times \prod_{j\neq i}\binom{y_{j,1}+y_{j,2}}{y_{j,1}}\rho_{j,1}^{y_{j,1}}\rho_{j,2}^{y_{j,2}},\end{aligned}$$

from which we deduce:

$$\frac{\bar{y}_{i,1}}{\rho_{i,1}} = \frac{\bar{y}_{i,2}}{\rho_{i,2}} = \pi(0)\sum_y (y_{i,1}+y_{i,2}+1)\Phi(y_1+y_2+e_i)\prod_j \binom{y_{j,1}+y_{j,2}}{y_{j,1}}\rho_{j,1}^{y_{j,1}}\rho_{j,2}^{y_{j,2}}.$$

The proof then follows from (17.2) and the equalities $\bar{x}_i = \bar{y}_{i,1}+\bar{y}_{i,2}$ and $\rho_i = \rho_{i,1}+\rho_{i,2}$.

Decomposing the flow size distribution into an arbitrary number of phases, we deduce similarly that the stationary distribution does not depend on the chosen Cox distribution and that all phases have the same flow throughput. Considering the limiting case where each phase is infinitely small, we conclude that the flow throughput of a class-*i* flow is equal to γ_i independently of its size. Equivalently, the mean per-bit delay of a class-i flow is equal to $1/\gamma_i$ independently of the considered bit in this data flow.

Sessions

We now assume that flows have i.i.d. sizes with exponential distribution but are generated within sessions. We first consider the simple case of two-flow sessions. Sessions of class-i flows arrive as a Poisson process of intensity λ_i, start with a flow of exponential size of mean $\sigma_{i,1}$ bits, which is followed by a flow of exponential size of mean $\sigma_{i,2}$ bits after a think-time of exponential duration. We let $\sigma_{i,1} + \sigma_{i,2} = \sigma_i$ so that the total traffic intensity generated by class-i flows is still equal to ρ_i.

It may again be easily verified that the associated queueing network is a Whittle network, where think-times are represented by infinite-server queues. The stationary distribution of the number of flows of each class is given by (17.3) under the stability condition (17.4). It is insensitive to the choice of $\sigma_{i,1}$ and $\sigma_{i,2}$ (provided $\sigma_{i,1} + \sigma_{i,2} = \sigma_i$) and to the mean think-time durations. It may also be shown as in Proposition 17.8.1 that the flow throughput is the same for the first flow and the second flow of the session.

These results extend to sessions with an arbitrary number of flows and Cox distributions for the flow sizes and the think-time durations. One may in fact represent virtually any traffic characteristics by considering as many types of sessions as necessary. The sizes and durations of successive flows and think-times within the same session may be correlated (e.g., each small flow is followed by a short think-time and the session ends with a large flow). The stationary distribution of the number of flows of each class is still given by (17.3) under the stability condition (17.4). Moreover, the flow throughput of a class-i flow is equal to γ_i independently of its size, the type of session it belongs to and its position in the session (e.g., first, second or last flow of the session). Equivalently, the mean per-bit delay of a class-i flow is equal to $1/\gamma_i$ independently of the considered bit within the flow and the considered flow within the session.

17.9 A single link

This section is devoted to the practically interesting case of a single link of capacity C bit/s. Flows have different rate limits and share the link capacity according to balanced fairness. In view of the above insensitivity results, we can restrict the analysis to the Markovian setting described in Section 17.3 where flows of each class arrive as a Poisson process and have i.i.d. sizes with exponential distribution.

No flow rate limit

We start with the simple case where flows do not have any individual rate limit. The model then corresponds to an $M/M/1$ queue of load ρ/C, where ρ denotes the traffic intensity in bit/s. The steady state distribution is:

$$\pi(x) = \pi(0)\left(\frac{\rho}{C}\right)^x.$$

We have:

$$\bar{x} = \frac{\rho}{C-\rho}$$

so that, in view of (17.2),

$$\gamma = C - \rho. \tag{17.1}$$

Thus the flow throughput is equal to the *residual capacity*, defined as the difference between link capacity and traffic intensity. This result may in fact be deduced from the following simple argument. Each flow gets all capacity not used by other flows. By ergodicity, the throughput of a flow of infinite size is equal to $C-\rho$ (since the capacity used by other flows is equal to the traffic intensity). The result (17.1) then follows from the fact that the mean throughput of a flow is independent of its size (cf. Section 17.8).

A common flow rate limit

Now assume that flows have a common rate limit $a = C/m$ for some positive integer m. The model then corresponds to an $M/M/m$ queue of load ρ/C. The steady state distribution is:

$$\pi(x) = \pi(0)\frac{\rho^x}{x!a^x} \text{ if } x \le m, \quad \pi(x) = \pi(m)\left(\frac{\rho}{C}\right)^{x-m} \text{ if } x > m.$$

As mentioned in Section 17.2, an interesting throughput performance metric is the steady state probability that flows do not get their rate limit, a. This is the probability S that the link is saturated, related to the Erlang C formula, which can be evaluated by means of the following simple recursive algorithm. Denote by $p(\cdot) = \pi(\cdot)/\pi(0)$ the unnormalized invariant measure. We have:

$$S = \frac{\bar{p}}{1 + p(1) + \ldots + p(m) + \bar{p}}$$

with

$$\bar{p} = \sum_{x>m} p(x).$$

The recursive algorithm is given by:

$$p(0) = 1, \quad p(x) = \frac{\rho}{xa}p(x-1) \quad \text{for } x = 1, \ldots, m,$$

and

$$\bar{p} = \frac{\rho}{C-\rho} p(m).$$

The flow throughput is related to the probability of saturation through the simple expression:

$$\gamma = a \frac{1-\rho}{1-\rho+S}.$$

Multiple rate limits

Finally, consider the general case where class-i flows have the rate limit $a_i > 0$. In view of (17.2), the balance function is defined by:

$$\Phi(x) = \prod_i \frac{1}{x_i! a_i^{x_i}} \quad \text{if } x.a \leq C,$$

and

$$\Phi(x) = \frac{1}{C} \sum_i \Phi(x - e_i) \quad \text{if } x.a > C.$$

We deduce from (17.3) the stationary distribution of the number of flows of each class:

$$\pi(x) = \pi(0) \prod_i \frac{1}{x_i!} \left(\frac{\rho_i}{a_i} \right)^{x_i} \quad \text{if } x.a \leq C, \tag{17.2}$$

and

$$\pi(x) = \frac{1}{C} \sum_i \rho_i \pi(x - e_i) \quad \text{if } x.a > C. \tag{17.3}$$

Probability of saturation

Assuming the link capacity and the flow rate limits are integers, the steady state probability S that the link is saturated can again be derived through a simple recursive algorithm. Let:

$$\forall n \in \mathbb{N}, \quad p(n) = \sum_{x:x.a=n} \frac{\pi(x)}{\pi(0)}.$$

Note that:

$$S = \frac{\bar{p}}{1 + p(1) + \ldots + p(C) + \bar{p}}.$$

with

$$\bar{p} = \sum_{n>C} p(n).$$

We denote by $\theta = \sum_i \rho_i$ the overall traffic intensity.

Property 17.9.1 *We have:*

$$p(0)=1, \quad p(n)=\sum_i \frac{\rho_i}{n} p(n-a_i) \quad \textit{for } n=1,\ldots,C,$$

with the convention that $p(n)=0$ *if* $n<0$, *and*

$$\bar{p}=\sum_i \frac{\rho_i \bar{p}_i}{C-\theta} \quad \textit{with} \quad \bar{p}_i=\sum_{n:C-a_i<n\le C} p(n).$$

Proof. The first part of the recursion follows from (17.2). We indeed have for all $n=1,\ldots,C$:

$$\begin{aligned}
p(n) &= \sum_{x:x.a=n} \frac{x.a}{n}\frac{\pi(x)}{\pi(0)}\\
&= \sum_{x:x.a=n}\sum_{i:x_i>0} \frac{\rho_i}{n}\frac{1}{(x_i-1)!}\left(\frac{\rho_i}{a_i}\right)^{x_i-1}\prod_{j\neq i}\frac{1}{x_j!}\left(\frac{\rho_j}{a_j}\right)^{x_j}\\
&= \sum_i \frac{\rho_i}{n}\sum_{x:\,(x+e_i).a=n}\frac{\pi(x)}{\pi(0)}\\
&= \sum_i \frac{\rho_i}{n} p(n-a_i).
\end{aligned}$$

Now using (17.3) we get:

$$\begin{aligned}
\bar{p} &= \sum_{x:x.a>C} \frac{\pi(x)}{\pi(0)}\\
&= \sum_{x:x.a>C} \frac{1}{C}\sum_i \rho_i \frac{\pi(x-e_i)}{\pi(0)}\\
&= \sum_i \frac{\rho_i}{C}\sum_{x:\,(x+e_i).a>C}\frac{\pi(x)}{\pi(0)}\\
&= \sum_i \frac{\rho_i}{C}(\bar{p}+\bar{p}_i).
\end{aligned}$$

from which the second part of the recursion easily follows.

Flow throughput

The flow throughput can be obtained by means of another recursive algorithm. Define for each class i:

$$\forall n\in\mathbb{N}, \quad q_i(n)=\sum_{x:x.a=n} x_i\frac{\pi(x)}{\pi(0)}.$$

In view of (17.2), the flow throughput of class i is given by:

$$\gamma_i=\rho_i\frac{1+p(1)+\ldots+p(m)+\bar{p}}{1+q_i(1)+\ldots+q_i(m)+\bar{q}_i}$$

with

$$\bar{q}_i = \sum_{n>C} q_i(n).$$

Property 17.9.2 *We have:*

$$q_i(0) = 0, \quad q_i(n) = \frac{\rho_i}{n} p(n-a_i) + \sum_j \frac{\rho_j}{n} q_i(n-a_j) \quad \text{for } n = 1,\ldots,C,$$

with the convention that $q_i(n) = 0$ *if* $n < 0$*, and*

$$\bar{q}_i = \rho_i \frac{\bar{p}_i + \bar{p}}{C-\rho} + \sum_j \frac{\rho_j \bar{q}_{ij}}{C-\rho} \quad \text{with} \quad \bar{q}_{ij} = \sum_{n:C-a_j<n\le C} q_i(n).$$

Proof. The proof is similar to that of Proposition 17.9.1. We have for all $n = 1,\ldots,C$:

$$\begin{aligned} q_i(n) &= \sum_{x:x.a=n} x_i \frac{x.a}{n} \frac{\pi(x)}{\pi(0)} \\ &= \sum_{x:x.a=n} x_i \sum_{j:x_j>0} \frac{\rho_j}{n} \frac{1}{(x_j-1)!} \left(\frac{\rho_j}{a_j}\right)^{x_j-1} \prod_{k\ne j} \frac{1}{x_k!} \left(\frac{\rho_k}{a_k}\right)^{x_k} \\ &= \sum_j \frac{\rho_j}{n} \sum_{x:\,(x+e_j).a=n} x_i \frac{\pi(x)}{\pi(0)} \\ &= \frac{\rho_i}{n} p(n-a_i) + \sum_j \frac{\rho_j}{n} q_i(n-a_j). \end{aligned}$$

The second part of the recursion follows from (17.3):

$$\begin{aligned} \bar{q}_i &= \sum_{x:x.a>C} x_i \frac{\pi(x)}{\pi(0)} \\ &= \sum_{x:x.a>C} \frac{x_i}{C} \sum_j \rho_j \frac{\pi(x-e_j)}{\pi(0)} \\ &= \sum_j \frac{\rho_j}{C} \sum_{x:\,(x+e_j).a>C} x_i \frac{\pi(x)}{\pi(0)} \\ &= \frac{\rho_i}{C}(\bar{p}_i + \bar{p}) + \sum_j \frac{\rho_j}{C}(\bar{q}_i + \bar{q}_{ij}). \end{aligned}$$

17.10 Performance bounds

We now provide explicit bounds on the flow throughput that prove useful for the performance evaluation of networks with several resources. For convenience, we focus on the per-bit delay τ_i of class i, which is defined as the inverse of the flow throughput γ_i of class i. We first assume that the capacity constraints reduce to (17.1). The impact of individual rate limits is described at the end of the section. We shall prove that, under balanced fairness,

$$\max_l \frac{A_{il}}{C_l - \theta_l} \leq \tau_i \leq \max_l \frac{A_{il}}{C_l} + \sum_l \frac{\theta_l}{C_l} \frac{A_{il}}{C_l - \theta_l}, \tag{17.1}$$

where θ denotes the line vector ρA.

Flows with a single capacity constraint

Before proving the inequalities (17.1), we consider the case where class-i flows are constrained by resource l only, in the sense that $A_{ir} = 0$ for all $r \neq l$. The bounds then coincide and we have:

$$\tau_i = \frac{A_{il}}{C_l - \theta_l}. \tag{17.2}$$

We give a direct proof of this result, which will be useful for the proof of (17.1). We need the following preliminary result:

Property 17.10.1 *Assume class-i flows are constrained by resource l only. Then resource l is saturated in any state x such that $x_i > 0$, that is*

$$\Phi(x) = \frac{1}{C_l} \sum_j A_{jl} \Phi(x - e_j). \tag{17.3}$$

Proof. The proof is by induction on the total number of flows $n \equiv |x|$. The property holds for $n = 1$ since, in view of (17.1),

$$\Phi(e_i) = \frac{A_{il}}{C_l}.$$

Now assume the property holds for $n = m$, for some $m \geq 1$. Let x be any state such that $x_i > 0$ and $n = m + 1$. By the induction hypothesis, we have for any resource r:

$$\sum_j \frac{A_{jl}}{C_l} \Phi(x - e_j) \geq \sum_{j,k} \frac{A_{jl}}{C_l} \frac{A_{kr}}{C_r} \Phi(x - e_j - e_k) = \sum_k \frac{A_{kr}}{C_r} \Phi(x - e_k).$$

Since the inequality holds for all r, it follows from (17.1) that:

$$\Phi(x) = \sum_j \frac{A_{jl}}{C_l} \Phi(x - e_j).$$

In view of (17.3), we get for all states x such that $x_i > 0$:

$$\pi(x) = \sum_j \frac{A_{jl}}{C_l} \rho_j \pi(x - e_j),$$

with the convention $\pi(x) = 0$ if $x \notin \mathbb{N}^N$. Thus,

$$\begin{aligned}
\bar{x}_i &= \sum_x x_i \pi(x) \\
&= \sum_j \frac{A_{jl}}{C_l} \rho_j \sum_x x_i \pi(x - e_j), \\
&= \sum_{j \neq i} \frac{A_{jl}}{C_l} \rho_j \sum_x x_i \pi(x - e_j) + \frac{A_{il}}{C_l} \rho_i \sum_x (x_i - 1)\pi(x - e_i) + \frac{A_{il}}{C_l} \rho_i \sum_x \pi(x - e_i), \\
&= \sum_{j \neq i} \frac{A_{jl}}{C_l} \rho_j \bar{x}_i + \frac{A_{il}}{C_l} \rho_i \bar{x}_i + \frac{A_{il}}{C_l} \rho_i, \\
&= \frac{\theta_l}{C_l} \bar{x}_i + \frac{A_{il}}{C_l} \rho_i.
\end{aligned}$$

Expression (17.2) then follows from (17.1).

Lower bound

The proof of the lower bound (17.1) is similar to the proof of (17.2). Replacing (17.3) by the following inequality, valid for all resources l in view of (17.1):

$$\Phi(x) \leq \frac{1}{C_l} \sum_j A_{jl} \Phi(x - e_j),$$

we obtain:

$$\tau_i \geq \frac{A_{il}}{C_l - \theta_l}.$$

Upper bound

To prove the upper bound, we add $L + 1$ "virtual" classes. Virtual class 0 has the same capacity constraint as class i and for all $l = 1, \ldots, L$, virtual class l uses resource l only and has the same requirement for this resource as class i. The original allocation vector ϕ and the new allocation vector $\tilde{\phi}$ associated with the virtual classes must satisfy the component-wise inequality:

$$\phi A + \tilde{\phi} \tilde{A} \leq C,$$

where by definition of the virtual classes, $\tilde{A}_{0l} = A_{il}$ and $\tilde{A}_{kl} = 1_{k=l} A_{il}$ for all $k, l = 1, \ldots, L$. We denote by $\tilde{x}$ the new network state associated with the virtual classes. The original network state is still denoted by x. The new balance function

$\tilde{\Phi}$ associated with balanced fairness is recursively defined by $\tilde{\Phi}(0) = 1$ and:

$$\tilde{\Phi}(x,\tilde{x}) = \max_l \left\{ \frac{A_{il}}{C_l} \tilde{\Phi}(x, \tilde{x} - e_0) + \frac{A_{il}}{C_l} \tilde{\Phi}(x, \tilde{x} - e_l) + \sum_j \frac{A_{jl}}{C_l} \tilde{\Phi}(x - e_j, \tilde{x}) \right\}, \quad (17.4)$$

with the convention $\tilde{\Phi}(x,\tilde{x}) = 0$ if $x \notin \mathbb{N}^N$ or $\tilde{x} \notin \mathbb{N}^{L+1}$.

We have the following key result:

Property 17.10.2 *For any state $x \in \mathbb{N}^N$,*

$$\tilde{\Phi}(x, e_0) + (\sum_l \frac{A_{il}}{C_l}) \tilde{\Phi}(x, 0) \leq (\max_l \frac{A_{il}}{C_l}) \tilde{\Phi}(x, 0) + \sum_l \tilde{\Phi}(x, e_l).$$

Proof. In view of Proposition 17.10.1, the inequality is equivalent to:

$$\tilde{\Phi}(x, e_0) \leq (\max_l \frac{A_{il}}{C_l}) \tilde{\Phi}(x, 0) + \sum_l \sum_j \frac{A_{jl}}{C_l} \tilde{\Phi}(x - e_j, e_l).$$

The proof is by induction on the total number of flows $n \equiv |x|$. The property holds for $n = 0$. Assume it holds for $n = m$ and let x be any state such that $n = m + 1$. We denote by r a resource that is saturated in state (x, e_0):

$$\tilde{\Phi}(x, e_0) = \frac{A_{ir}}{C_r} \tilde{\Phi}(x, 0) + \sum_k \frac{A_{kr}}{C_r} \tilde{\Phi}(x - e_k, e_0).$$

By the induction hypothesis, we have:

$$\tilde{\Phi}(x, e_0) \leq \frac{A_{ir}}{C_r} \tilde{\Phi}(x, 0) + (\max_l \frac{A_{il}}{C_l}) \sum_k \frac{A_{kr}}{C_r} \tilde{\Phi}(x - e_k, 0) + \sum_l \sum_{j,k} \frac{A_{kr}}{C_r} \frac{A_{jl}}{C_l} \tilde{\Phi}(x - e_j - e_k, e_l).$$

Now it follows from (17.4) that for any class j:

$$\forall l \neq r, \quad \sum_k \frac{A_{kr}}{C_r} \tilde{\Phi}(x - e_j - e_k, e_l) \leq \tilde{\Phi}(x - e_j, e_l),$$

and

$$\frac{A_{ir}}{C_r} \tilde{\Phi}(x - e_j, 0) + \sum_k \frac{A_{kr}}{C_r} \tilde{\Phi}(x - e_j - e_k, e_r) \leq \tilde{\Phi}(x - e_j, e_r).$$

We deduce that:

$$\begin{aligned} \tilde{\Phi}(x, e_0) \leq\ & \frac{A_{ir}}{C_r} \tilde{\Phi}(x, 0) + (\max_l \frac{A_{il}}{C_l}) \sum_k \frac{A_{kr}}{C_r} \tilde{\Phi}(x - e_k, 0) \\ & + \sum_l \sum_j \frac{A_{jl}}{C_l} \tilde{\Phi}(x - e_j, e_l) - \sum_j \frac{A_{jr}}{C_r} \frac{A_{ir}}{C_r} \tilde{\Phi}(x - e_j, 0). \end{aligned}$$

Thus the proof will be completed if we show that:

$$\frac{A_{ir}}{C_r}\tilde{\Phi}(x,0)+(\max_l \frac{A_{il}}{C_l})\sum_k \frac{A_{kr}}{C_r}\tilde{\Phi}(x-e_k,0)-\sum_j \frac{A_{jr}}{C_r}\frac{A_{ir}}{C_r}\tilde{\Phi}(x-e_j,0)\le(\max_l \frac{A_{il}}{C_l})\tilde{\Phi}(x,0).$$

But this inequality may also be written:

$$\left(\max_l \frac{A_{il}}{C_l}-\frac{A_{ir}}{C_r}\right)\left(\tilde{\Phi}(x,0)-\sum_j \frac{A_{jr}}{C_r}\tilde{\Phi}(x-e_j,0)\right)\ge 0,$$

which is satisfied in view of (17.4).

Under the stability condition (17.3), there exists an $(L+1)$-dimensional vector $\tilde{\rho}$ such that the following component-wise strict inequality is satisfied:

$$\tilde{\theta}\equiv\rho A+\tilde{\rho}\tilde{A}<C.$$

Let $\tilde{\tau}_l$ be the corresponding per-bit delay of virtual class l, for all $l=0,1,\ldots,L$. To prove the upper bound, we use the fact that:

$$\lim_{\tilde{\rho}\to 0}\tilde{\tau}_0=\tau_i$$

and, in view of (17.2),

$$\forall l=1,\ldots,L,\quad \tilde{\tau}_l=\frac{A_{il}}{C_l-\tilde{\theta}_l}.$$

In particular,

$$\forall l=1,\ldots,L,\quad \lim_{\tilde{\rho}\to 0}\tilde{\tau}_l=\frac{A_{il}}{C_l-\theta_l}.$$

Now it follows from (17.1) and (17.3) that:

$$\forall l=0,1,\ldots,L,\quad \lim_{\tilde{\rho}\to 0}\tilde{\tau}_l=\frac{\sum_x \tilde{\Phi}(x,e_l)\rho^x}{\sum_x \tilde{\Phi}(x,0)\rho^x}.$$

Using Proposition 17.10.2, we obtain:

$$\tau_i+\sum_l \frac{A_{il}}{C_l}\le\max_l \frac{A_{il}}{C_l}+\sum_l \frac{A_{il}}{C_l-\theta_l},$$

from which the upper bound (17.1) directly follows.

Flow rate limits.

In the presence of per-flow rate constraints (17.2), the bounds become:

$$\max\left\{\frac{1}{a_i},\max_l \frac{A_{il}}{C_l-\theta_l}\right\}\le\tau_i\le\max\left\{\frac{1}{a_i},\max_l \frac{A_{il}}{C_l}\right\}+\sum_l \frac{\theta_l}{C_l}\frac{A_{il}}{C_l-\theta_l}. \qquad (17.5)$$

The proof is similar and omitted here.

17.11 Examples

This section presents numerical results for the various networks introduced in Section 17.2. The flow throughput under balanced fairness is compared with the conservative bound derived in Section 17.10 and with the flow throughput under max-min fairness and proportional fairness. The latter is obtained by simulation in the Markovian setting described in Section 17.3 (except in the specific cases where these allocations coincide with balanced fairness). Each simulation point corresponds to expression (17.2) where the mean number of flows is evaluated over 1,000,000 events after a warm-up period of 100,000 events. All expressions and bounds concern implicitly balanced fairness.

Wireline networks

Consider the 2-link linear network of Figure 17.1 with capacity constraints:

$$\phi_1 + \phi_3 \leq 1, \quad \phi_2 + \phi_3 \leq 1.$$

The stability condition is $\theta_1 < 1$ and $\theta_2 < 1$, where $\theta_1 = \rho_1 + \rho_3$ and $\theta_2 = \rho_2 + \rho_3$ are the traffic intensities at the first and the second link. Class-1 and class-2 flows are constrained by a single resource so that, in view of (17.2),

$$\gamma_1 = 1 - \theta_1, \quad \gamma_2 = 1 - \theta_2.$$

For class-3 flows, it follows from (17.1) that:

$$\gamma_3 \geq \frac{(1-\theta_1)(1-\theta_2)}{1-\theta_1\theta_2}.$$

The left-hand graph of Figure 17.10 illustrates the tightness of this bound for equal traffic intensities $\rho_1 = \rho_2 = \rho_3$. The bound is compared to the exact expression when the total traffic intensity at each link $\theta_1 = \theta_2$ varies from 0 to 1. As mentioned in Section 17.7, proportional fairness coincides with balanced fairness in this particular case. The right-hand graph of Figure 17.10 shows that max-min fairness gives very similar results.

Now consider the tree network of Figure 17.2 with three unit capacity branches and a common root link of capacity 2. The capacity constraints are:

$$\phi_1 + \phi_2 + \phi_3 \leq 2, \quad \phi_1 \leq 1, \ \phi_2 \leq 1, \ \phi_3 \leq 1.$$

For equal traffic intensities, the stability condition is $\theta < 2$ where $\theta = \rho_1 + \rho_2 + \rho_3$ denotes the traffic intensity at the common root link. It follows from (17.1) that:

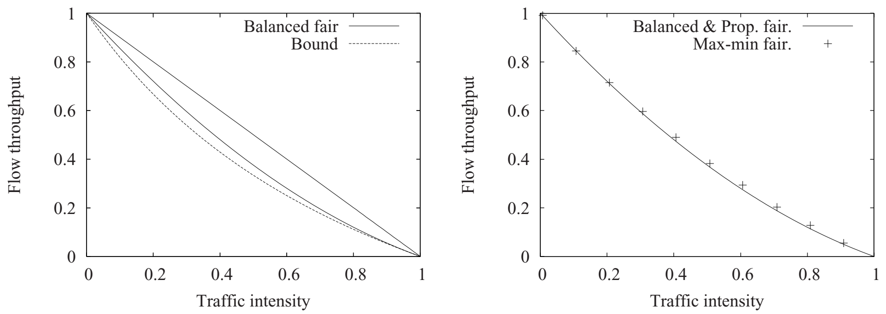


Fig. 17.10: Flow throughput of class 3 in the linear network of Figure 17.1.

$$\gamma_1 = \gamma_2 = \gamma_3 \geq \frac{2(2-\theta)(3-\theta)}{12-3\theta-\theta^2}.$$

The results are shown in Figure 17.11 with respect to θ. Again, the throughput performance of proportional fairness and max-min fairness is very similar to that of balanced fairness.

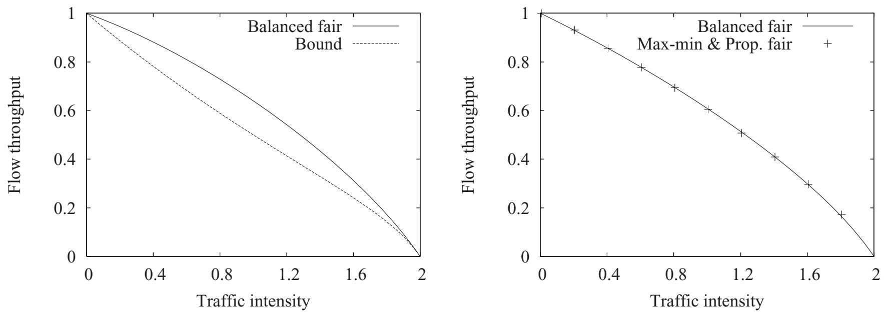


Fig. 17.11: Flow throughput in the tree network of Figure 17.2.

Traffic splitting

Consider the network of Figure 17.3 with capacity constraints:

$$\phi_1 \leq 1/2, \quad \phi_1 + \phi_2 \leq 1, \quad \phi_2 + \phi_3 \leq 1.$$

The stability condition is $\theta_1 < 1/2$, $\theta_2 < 1$, $\theta_3 < 1$ with $\theta_1 = \rho_1$, $\theta_2 = \rho_1 + \rho_2$ and $\theta_3 = \rho_2 + \rho_3$. It follows from (17.1) that:

$$\gamma_1 \geq \frac{(1/2-\theta_1)(1-\theta_2)}{1-\theta_1\theta_2-\theta_2/2}, \quad \gamma_2 \geq \frac{(1-\theta_2)(1-\theta_3)}{1-\theta_2\theta_3}, \quad \gamma_3 = 1-\theta_3.$$

For equal traffic intensities, the stability condition reduces to $\theta < 3/2$ where $\theta = \rho_1+\rho_2+\rho_3$ denotes the total traffic intensity. The results are shown in Figure 17.12 with respect to θ.

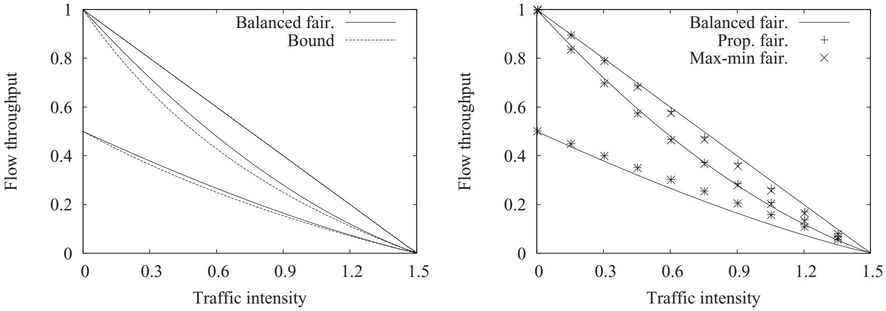


Fig. 17.12: Flow throughput in the network of Figure 17.3 (classes 1,2,3, from bottom to top).

Wireless networks

Now consider the wireless access point of Figure 17.4, characterized by the unique capacity constraint:

$$\frac{\phi_1}{c_1}+\frac{\phi_2}{c_2}+\frac{\phi_3}{c_3} \leq 1.$$

The stability condition is $\theta < 1$, where θ is the system load:

$$\theta = \frac{\rho_1}{c_1}+\frac{\rho_2}{c_2}+\frac{\rho_3}{c_3}.$$

Proportional fairness gives for each class $i = 1,2,3$:

$$\forall x \neq 0, \quad \phi_i(x) = \frac{x_i}{\sum_i x_i} c_i.$$

The allocation satisfies the balance property (17.1) and coincides with balanced fairness. Under these allocations, the access point serves each flow the same fraction of time, so that the transmission rate of each flow is proportional to its coding rate. In view of (17.2), the flow throughput of each class $i = 1,2,3$ is given by:

$$\gamma_i = c_i(1-\theta).$$

Max-min fairness, on the other hand, gives for each class $i = 1,2,3$:

$$\forall x \neq 0, \quad \phi_i(x) = \frac{x_i}{\sum_i x_i/c_i}.$$

Thus the transmission rate is the same for all flows. This results in a *discriminatory* allocation of the radio resource: the access point serves each flow a fraction of time that is inversely proportional to its coding rate. The resulting flow throughput is shown in Figure 17.13 for $c_1 = 5$, $c_2 = 1$, $c_3 = 1/2$ and equal traffic intensities $\rho_1 = \rho_2 = \rho_3$. The stability condition $\theta < 1$ imposes that the total traffic intensity is less than $15/16$. We observe that class-1 flows are strongly penalized by max-min fairness. Since these flows contribute to a small fraction of the overall system load θ, the benefit for other classes is marginal. We conclude that the radio resource should not be allocated according to max-min fairness for this particular system.

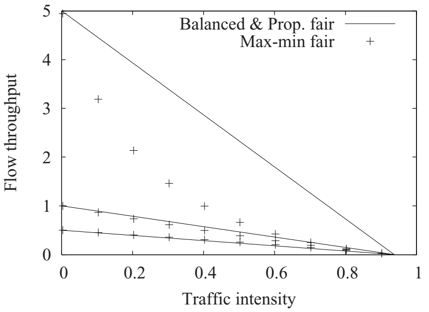


Fig. 17.13: Flow throughput at a wireless access point (classes 1,2,3, from top to bottom).

Now consider the additional constraint of a wireline backhaul link of c bit/s, as shown in Figure 17.5. The stability condition becomes $\theta_1 < 1$ and $\theta_2 < c$, where θ_1 and θ_2 correspond to the load of the wireless link and to the traffic intensity on the wireline link, respectively:

$$\theta_1 = \frac{\rho_1}{c_1} + \frac{\rho_2}{c_2} + \frac{\rho_3}{c_3}, \quad \theta_2 = \rho_1 + \rho_2 + \rho_3.$$

In view of (17.1), we get for each class $i = 1,2,3$:

$$\gamma_i \geq \left(\max\left\{\frac{1}{c_i}, \frac{1}{c}\right\} + \frac{\theta_1}{c_i(1-\theta_1)} + \frac{\theta_2}{c(c-\theta_2)} \right)^{-1}.$$

The results are shown in Figure 17.14 for the same parameters as above and $c = 2$. The inequality $\theta_1 < 1$ is more restrictive than $\theta_2 < 2$ in this case, so that the system is again stable if and only if the total traffic intensity is less than $15/16$. We observe that balanced fairness provides a good approximation to proportional fairness, whose throughput performance is much better than that of max-min fairness.

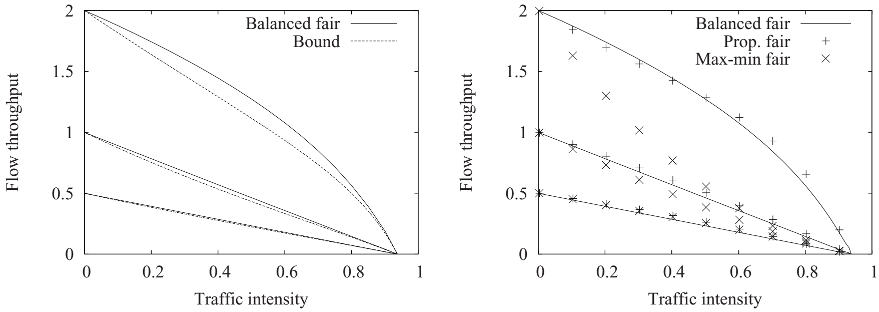


Fig. 17.14: Flow throughput at a wireless access point with wireline backhaul (classes 1,2,3, from top to bottom).

Ad-hoc networks

The ad-hoc network of Figure 17.7 is characterized by the following capacity constraints:

$$2\phi_1+2\phi_2+3\phi_3 \le 1, \quad 3\phi_1+\phi_2+2\phi_3 \le 1, \quad \phi_1+3\phi_2+2\phi_3 \le 1.$$

The stability condition is given by $\theta_1 < 1$, $\theta_2 < 1$, $\theta_3 < 1$ with

$$\theta_1 = 2\rho_1+2\rho_2+3\rho_3, \quad \theta_2 = 3\rho_1+\rho_2+2\rho_3, \quad \theta_3 = \rho_1+3\rho_2+2\rho_3.$$

It follows from (17.1) that:

$$\gamma_1 \ge \left(3+\frac{2\theta_1}{1-\theta_1}+\frac{3\theta_2}{1-\theta_2}+\frac{\theta_3}{1-\theta_3}\right)^{-1}.$$

Figure 17.15 shows class-1 flow throughput with respect to class-1 traffic intensity ρ_1 for equal traffic intensities. Note that the stability condition is given by $\rho_1 < 3/7$ in this case.

Flow rate limits

Finally, we consider a wireline link of C bit/s shared by flows with $N = 3$ different rate limits, $a_1, a_2, a_3 < C$. The stability condition is given by $\theta < C$ where $\theta =$

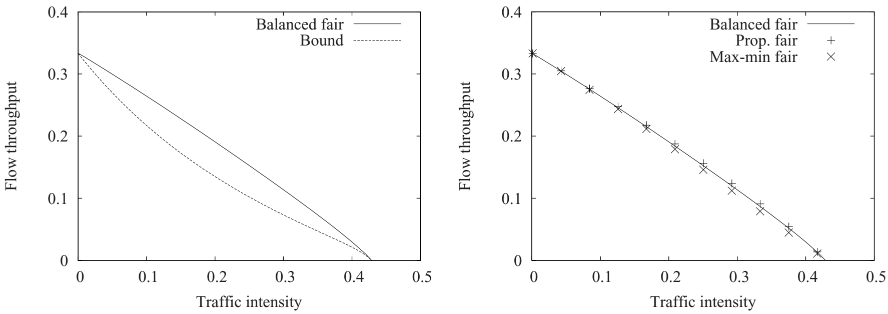


Fig. 17.15: Class-1 flow throughput in the ad-hoc network of Figure 17.7.

$\rho_1 + \rho_2 + \rho_3$ denotes the total traffic intensity. It follows from (17.5) that for each class $i = 1,2,3$:

$$\gamma_i \geq \left(\frac{1}{a_i} + \frac{\theta}{C(C-\theta)} \right)^{-1}.$$

Figure 17.16 gives the corresponding results with respect to θ for $C = 10$, $a_1 = 2$, $a_2 = 1$, $a_3 = 1/2$ and equal traffic intensities $\rho_1 = \rho_2 = \rho_3$. Proportional fairness and max-min fairness coincide in this case.

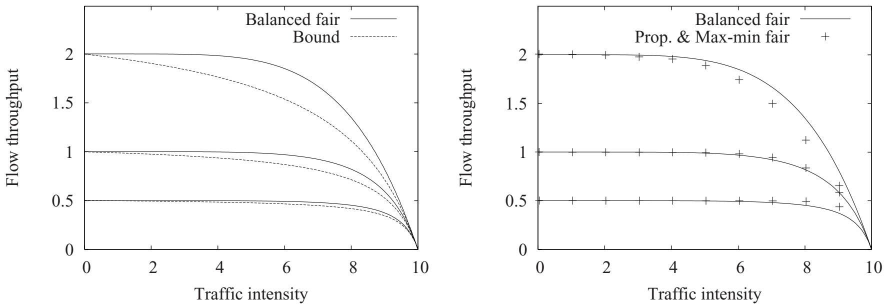


Fig. 17.16: Flow throughput for a wireline link with different flow rate limits (classes 1,2,3, from top to bottom).

17.12 Open issues

While balanced fairness closely approximates proportional fairness in all considered examples, there is no theoretical result that supports this evidence except for some structural properties obtained by Massoulié [15]. It seems even more difficult to assess the throughput performance of max-min fairness which differs significantly from that of balanced fairness in some cases. For both proportional fair and max-min fair allocations, deriving bounds or approximations that only depend on the capacity region (17.1) and on the vector of traffic intensities is a very challenging task.

There has been very little work on non-convex capacity regions. First, the optimization problem that defines proportional fairness does not necessarily have a unique solution. Next, the stability condition is unknown in general. In particular, the network may be stable even though the vector of traffic intensities does not belong to the capacity region but to the convex hull of this capacity region.

While the focus has been on data traffic only, data transfers must often share network resources with flows of other applications like the telephone or audio and video streaming. These flows have packet delay constraints that require specific rate adaptation and scheduling algorithms. The impact of this traffic and its particular control schemes on the throughput of data flows is not very well understood. Explicit results can be obtained by the so-called quasi-stationary approach, where the time-scale of data flows is assumed to be very different from that of other flows. It remains to determine the conditions under which these results provide bounds or tight approximations for the exact throughput performance.

17.13 Bibliographical notes

The modeling of data links as processor-sharing queues started with the analysis of wireless systems by Telatar and Gallager [22] and Stamatelos and Koukoulidis [21]. Based on the observation that the transmission control protocol, TCP, shares resources in an approximately fair way, Heyman, Lakshman and Neidhardt [8] and Massoulié and Roberts [16] applied similar models to wireline networks. Practical dimensioning rules were developed on this basis by Berger and Kogan [2]. Ben Fredj et al. observed the insensitivity of the results to detailed traffic characteristics like the structure of user sessions [1]. Many papers proposed modified models that account more precisely for the way bandwidth is shared by TCP, see [12] for instance.

The notion of max-min fairness was introduced for communication networks by Bertsekas and Gallager [3]. Kelly and his coauthors [10, 17] introduced the notion of proportional fairness and identified a class of decentralized algorithms that realize this allocation. Various extensions of these results were obtained by Low and Lapsley [14], Mo and Walrand [18], Massoulié and Roberts [17] and others.

Stability issues were addressed by De Veciana, Lee and Konstantopoulos [23], Bonald and Massoulié [4], Ye [24] and Lin, Shroff and Srikant [13] for various allocations, including max-min fairness and proportional fairness. The first analytical performance result for networks with several links was derived by Massoulié and Roberts [16]. The notion of balanced fairness was introduced for wireline networks by Bonald and Proutière [6] and generalized to any network with a convex capacity region by Bonald, Massoulié, Proutière and Virtamo [5]. These papers also characterize the networks for which max-min fairness and proportional fairness satisfy the balance property. Finally, the recursion derived by Bonald and Virtamo [7] for multirate systems is the analogue of that derived by Kaufman [9] and Roberts [19] for circuit-switched networks.

Appendix

We recall the definition and stationary distribution of Whittle networks. For details, we refer the reader to the book by Serfozo [20].

Consider a network of N queues. External customer arrivals at queue i form a Poisson process of intensity ν_i, with $\sum_i \nu_i > 0$. After service completion at queue i, a customer is routed to queue j with probability p_{ij} and leave the network with probability $1 - \sum_j p_{ij}$. All customers eventually leave the network so that the arrival rate λ_i at queue i is uniquely defined by the traffic equations:

$$\lambda_i = \nu_i + \sum_j \lambda_j p_{ji}, \quad i = 1, \ldots, N.$$

The service requirements are independent, exponentially distributed of mean σ_i at queue i. We denote by $\rho_i = \lambda_i \sigma_i$ the traffic intensity at queue i.

We denote by x_i the number of customers present at queue i and by x the corresponding line vector. The service rate of queue i is a function ϕ_i of the network state x, with $\phi_i(x) = 0$ if and only if $x_i = 0$. We say that the network is a Whittle network if the following balance property is satisfied:

$$\forall i, j, \quad \forall x, \quad \phi_i(x)\phi_j(x - e_i) = \phi_j(x)\phi_i(x - e_j),$$

where we use the convention that $\phi(x) = 0$ if $x \notin \mathbb{N}^N$. This is equivalent to the existence of a balance function Φ such that $\Phi(0) = 1$ and:

$$\forall x \neq 0, \quad \Phi(x) = \frac{1}{\phi_{i_1}(x)\phi_{i_2}(x - e_{i_1}) \ldots \phi_{i_n}(e_{i_n})},$$

where $x, x - e_{i_1}, x - e_{i_1} - e_{i_2}, \ldots, e_{i_n}, 0$ denotes any direct path from state x to state 0. The stationary distribution of the network state is then given by:

$$\forall x, \quad \pi(x) = \pi(0)\Phi(x)\rho^x,$$

under the stability condition:

$$\sum_x \Phi(x)\rho^x < \infty,$$

where we use the notation:

$$\rho^x \equiv \prod_i \rho_i^{x_i}.$$

References

1. S. Ben Fredj, T. Bonald, A. Proutière, G. Regnié and J.W. Roberts, Statistical bandwidth sharing: A study of congestion at flow level, in: *Proc of ACM SIGCOMM*, 2001.
2. A.W. Berger, Y. Kogan, Dimensioning bandwidth for elastic traffic in high-speed data networks, IEEE/ACM Trans. on Networking 8(5) (2000) 643–654.
3. D. Bertsekas and R. Gallager, *Data Networks*, Prentice Hall, 1987.
4. T. Bonald and L. Massoulié, Impact of fairness on Internet performance, in: *Proc. of ACM SIGMETRICS/ Performance*, 2001.
5. T. Bonald, L. Massoulié, A. Proutière, J. Virtamo, A queueing analysis of max-min fairness, proportional fairness and balanced fairness, Queueing Systems 53 (2006) 65–84.
6. T. Bonald and A. Proutière, Insensitive bandwidth sharing in data networks, Queueing Systems 44(1) (2003) 69–100.
7. T. Bonald and J. Virtamo, A recursive formula for multirate systems with elastic traffic, IEEE Communications Letters 9 (2005) 753–755.
8. D.P. Heyman, T.V. Lakshman, A.L. Neidhardt, A new method for analysing feedback-based protocols with applications to engineering Web traffic over the Internet, in: *Proc. of ACM SIGMETRICS*, 1997.
9. J. S. Kaufman, Blocking in a shared resource environment, IEEE Trans. Commun. 29 (1981) 1474–1481.
10. F.P. Kelly, Charging and rate control for elastic traffic, European Transactions on Telecommunications 8 (1997) 33–37.
11. F.P. Kelly, A. Maulloo and D. Tan, Rate control for communication networks: Shadow prices, proportional fairness and stability, Journal of the Operat. Res. Society 49 (1998).
12. A. Kherani and A. Kumar, Stochastic models for throughput analysis of randomly arriving elastic flows in the Internet, in: *Proc. of IEEE INFOCOM*, 2002.
13. X. Lin, N.B. Shroff, R. Srikant, On the connection-level stability of congestion-controlled communication networks, IEEE Trans. on Information Theory 54(5) (2008) 2317–2338.
14. S.H. Low and D.E. Lapsley, Optimization flow control, I: Basic algorithm and convergence, IEEE/ACM Trans. on Networking 7(6) (1999) 861–875.
15. L. Massoulié, Structural properties of proportional fairness: stability and insensitivity, Annals of Applied Probability 17(3) (2007) 809–839.
16. L. Massoulié and J.W. Roberts, Bandwidth sharing and admission control for elastic traffic, Telecommunication Systems 15 (2000) 185–201.
17. L. Massoulié and J.W. Roberts, Bandwidth sharing: Objectives and algorithms, IEEE/ACM Trans. on Networking 10(3) (2002) 320–328.
18. J. Mo and J. Walrand, Fair end-to-end window-based congestion control, IEEE/ACM Trans. on Networking 8(5) (2000) 556–567.
19. J. W. Roberts, A service system with heterogeneous user requirement, in *Performance of Data Communications Systems and Their Applications*, G. Pujolle, Ed. Amsterdam, The Netherlands: North-Holland, 1981, pp. 423–431.
20. R.F. Serfozo, *Introduction to Stochastic Networks*, Springer Verlag, 1999.
21. G.M. Stamatelos and V.N. Koukoulidis, Reservation-based bandwidth allocation in a radio ATM network, IEEE/ACM Trans. on Networking 5(3) (1997) 420–428.

22. I.E. Telatar and R.G. Gallager, Combining queueing theory with information theory for multiaccess, IEEE Journal Selected Areas in Comm. 13 (1995) 963–969.
23. G. de Veciana, T.J. Lee and T. Konstantopoulos, Stability and performance analysis of networks supporting elastic services, IEEE/ACM Trans. on Networking 9(1) (2001) 2–14.
24. H.Q. Ye, Stability of data networks under an optimization-based bandwidth allocation, IEEE Trans. Aut. Control 48(7) (2003) 1238–1242.
25. S. Zachary, A note on insensitivity in stochastic networks, Journal of Applied Probability 44(1) (2007) 238–248.

Chapter 18
Modeling a Hospital Queueing Network

Stefan Creemers and Marc Lambrecht

Abstract Healthcare systems differ intrinsically from manufacturing systems. As such, they require a distinct modeling approach. In this article, we show how to construct a queueing network of a general class of healthcare systems. In order to analyze such networks, we use the parametric decomposition approach. Using this approach the network is decomposed into a set of single queueing systems which can be analyzed separately. Afterwards, results of these single queueing systems can be aggregated and general performance measures of the queueing network are obtained. In addition, we develop new expressions to assess the impact of service outages and use the queueing network to approximate patient flow times and to evaluate a number of practical applications.

18.1 Introduction

Whereas the origin of queueing theory dates back from the beginning of the previous century, networks of queues have only been studied for a few decades. The pioneering works of Jackson (1957 and 1963) showed that the stationary distribution of the number of customers in queue at a queueing network, is a product form of the stationary distributions at the individual workstations of the network. As a consequence, a queueing network can be decomposed into separate building blocks (i.e. the individual workstations) that can be analyzed separately to obtain the solution to the network as a whole. This approach is referred to as the parametric decomposition approach. The main advantage of the approach is that it enables the study of, otherwise intractable, complex queueing networks.

Stefan Creemers and Marc Lambrecht
Faculty of Business and Economics, Department of Decision Sciences and Information Management, Catholic University of Leuven , Naamsestraat 69, 3000 Leuven, Belgium, e-mail: firstname.lastname@econ.kuleuven.be

R.J. Boucherie and N.M. van Dijk (eds.), *Queueing Networks: A Fundamental Approach*,
International Series in Operations Research & Management Science 154,
DOI 10.1007/978-1-4419-6472-4_18, © Springer Science+Business Media, LLC 2011

Unfortunately, the results obtained by Jackson (1957 and 1963) are only valid in so-called "Jackson networks" (i.e. queueing networks which assume Poisson arrival and service processes). When assuming a generalized queueing network (featuring general service and arrival processes), the product form solution no longer holds. As such, one requires another means to "link" the separate building blocks of the queueing network. This link is established in the form of a "linking equation". More specifically, a linking equation approximates the stochastic nature of the outgoing stream of customers at one of the workstations of the network. Using this information, we can assess the stochastic nature of the inflow of customers at the workstations further down the queueing network. As such, a linking equation literally "links" the results obtained at the separate workstations to obtain the solution of the network as a whole. Marshall (1968) was the first to study the stochastic nature of the outflow of customers at a queueing workstation. Ever since, a wide variety of linking equations (applicable to a wide variety of settings) has been developed. We refer the reader to Shanthikumar and Buzacott (1981), Buzacott and Shanthikumar (1985), Bitran and Tirupati (1988) and Suri, Sanders and Kamath (1993) for a nice review.

Among others, these results have been extended and implemented in Whitt's Queueing Network Analyzer (1983), a powerfull tool that allows the analysis of a wide variety of complex queueing networks. Other noteworthy contributions to the domain of parametric decomposition of queueing networks include the works of Whitt (1994, 1995 and 1999a), Bitran and Tirupati (1988) and Lambrecht, Ivens and Vandaele (1998). A comprehensive overview of research on queueing networks in general and the parametric decomposition method in particular may be found with Askin (1993) and Hopp and Spearman (2000).

Queueing networks however, have mainly been studied in a manufacturing setting. Applications towards services in general and healthcare in particular are rarely seen. One of the reasons thereof is the difficulty of implementing the peculiarities of a service system into a methodology that is focussed on manufacturing systems. In what follows we discuss which problems may arise when modeling complex hospital queueing networks. Next we demonstrate how to use the parametric decomposition approach to model such queueing networks. In addition, we develop new expressions to assess the impact of service outages in a healthcare setting. The queueing network is used to test a variety of practical problems. More specifically, we demonstrate the impact on system performance resulting from the reduction of service outages and illustrate the beneficial effects of pooling. Moreover, we develop an optimization model that enables us to determine the optimal number of patients to be treated during a service session (e.g. a consultation time block). Finally we present some conclusions.

18.2 Problem Description

An important feature of healthcare processes (or services in general) is that the demand for resources is to a large extent unscheduled. As a consequence, there is a permanent mismatch between the demand for a treatment and the available capacity. Moreover, timely care is very important so interrupts are common in healthcare processes (the sense of urgency is almost always present). No wonder that healthcare is riddled with delays. No need to come up with a convincing example, we have all experienced that phenomenon. Delays are highly undesirable, not only from a psychological point of view (patient satisfaction) but also from an economic point of view. Government reimbursement systems are more and more based on a Justified Length of Stay (JLoS) system. DRG's (Diagnosis Related Groups) are characterized by a minimum and maximum length of stay (depending on parameters such as severity of the illness, age of the patient, . . .). If a patient is dismissed before the JLoS is over, the hospital still collects a full reimbursement. On the other hand, if the patient remains in care for a period which exceeds the limit of the JLoS, the hospital has to pay for the extra costs involved. The JLoS of a DRG is determined in function of a national average length of stay. The system stimulates hospitals to continuously improve their performance. Moreover, improper scheduling and malfunctioning logistical systems cause lengths of stay that are too long. Insurance companies may reject reimbursement of these "denied days" because the delay is not medically necessary Hall, Belson, Muralli and Dessouky (2006). Delays also create a "hidden" hospital in analogy with the hidden company. In other words, such a hospital creates wasteful overhead.

Hall (2006) coined the term patient flow. It represents the ability of the healthcare system to serve patients quickly, reliably and efficiently as they move through stages of care. Queue and delay analysis can produce dramatic improvements in medical performance, patient satisfaction and cost efficiency of healthcare. Healthcare systems can be represented as a complex queueing network. The queueing models are helpful to determine the capacity levels (and the allocation of capacity) needed to respond to demands in a timely fashion (minimizing the delay). There is a demand side (the patient mix and the associated variability in the arrival stream) and a supply side (the hospital resources such as surgeons, nurses, operating rooms, waiting rooms, recovery, imaging machines, laboratories) in any healthcare process. Moreover, both demand and supply are inherently stochastic. This stochastic nature creates disturbances and outages during the process. It is the combination of capacity analysis and variability that makes queueing theory so attractive. The major objective is to identify factors influencing the flow time of patients, to identify levers of improvement and to analyze trade-offs. In this article we try to address some of the issues mentioned above.

Queueing models have been applied in numerous industrial settings and service industries. The number of applications in healthcare, however, is relatively small. This is probably due to a number of unique healthcare related features that make queueing problems particularly difficult to solve. In this section, we will review these features and where appropriate we will shortly discuss the methodological impact.

Before we dig into this issue, let's first discuss two important modeling issues in healthcare: the performance measures and the issue of pooled capacity.

The performance measures in healthcare systems focus on internal and external delays. The internal delay refers to the sojourn time of patients inside the hospital before treatment. The external delay refers to the phenomenon of waiting lists. Manufacturing systems may buffer with finished goods inventory, service systems rely more on time buffers and capacity buffers. Another important performance measure is related to the target occupancy (utilization) levels of resources. Average occupancy targets are often preferred by government and other institutional agents. Hereby, higher occupancy levels are preferred, but this results in longer delays. We are often confronted with conflicting objectives. Instead of determining capacity needs based on (target) occupancy levels, it is preferable to focus on delays. The key issue in delay has to do with the tail probability of the waiting time. The tail probability refers to the probability that a patient has to wait more than a specified time interval. Capacity needs (e.g. staffing) of an emergency department should be based on an upper bound on the fraction of patients who experience a delay of more than a specific time interval before receiving care from a physician (Green and Soares, 2007). The second modeling issue has to do with pooling. In general, pooling refers to the phenomenon that available inventory or capacity is shared among various sources of demand (well known examples are location pooling, commonality or flexible capacity). Pooling is based on the principle of aggregation and mostly comes down to the fact that we can handle uncertainty with less inventory or capacity. In healthcare systems, resources are usually dedicated to specific patient types, hospitals have separate units or departments by diagnostic type and bed flexibility is almost non-existing. As a result, pooling is absent. This explains the fact that most queueing models reported in the literature are dealing with parts of the hospital. Queueing models, however, can be used to model hospital wide systems and to evaluate the benefits of greater versus less specialization of care units or other resources (scanners, labs, . . .).

Let's now turn to a number of unique healthcare related features making queueing models in healthcare difficult to model and to solve.

Re-entry of patients and stochastic routings

During consultation, patients may be routed to different facilities. The routing of a patient through hospital facilities is not deterministic. Instead, during the diagnosis stage there is a probabilistic routing. Moreover, patients require in many cases several consultations before surgery. Even after a patient is discharged from the hospital after surgery and recovery, the patient is subjected to a number of follow-up consultations. In other words, the queueing model must take care of re-entry of patients, creating additional work on top of the new patients. In most cases, the re-entry is correlated.

Service sessions for consultation and surgery

In most queueing models time is considered as continuous and events are spread out over this continuous time scale. In services in general and in healthcare more specifically, resources are not continuously available. Instead, time is divided into "service sessions" for consultation (e.g. twice a week) or surgery (e.g. one day per week). Consequently we have to focus on service processes in which service takes place during predefined service sessions. Vacation models observe the queueing behavior of such systems in which servers are available during certain time intervals and are on "vacation" during the other time intervals.

Capacity related issues

Hospitals operate within strict business restrictions. Resources are usually very scarce and consequently hospitals operate under high capacity utilization conditions. The so-called heavy traffic conditions are present. Heavy traffic conditions assume that all stations in the network are critically loaded. In such an environment, inaccurate results have a large impact on resulting performance measures.

Modeling of absences, disturbances and interruptions

An important determinant of the flow time is variability. We distinguish two types of variability. Natural variability is variability that is inherent to the system process. Natural variability is much more substantial in healthcare as compared to manufacturing environments. Second, we have variability that can be related or assigned to a specific external cause. This variability is caused by unplanned absences of medical staff or interruptions during service operations. It is well known that variability induces waiting time. As a result the time available during consultation is often exceeded. This in turn is remedied by allowing overtime. Unfortunately, overtime modeling is a non-trivial issue in queueing.

18.3 A hospital queueing system

The features discussed in the previous section considerably complicate the modeling exercise. In order to demonstrate how to implement the features in a queueing model, we use an example hospital queueing system. The example concerns a typical hospital department involving consultation, surgery and recovery. The example we use throughout this paper is inspired by a real life case of the orthopedic department of the Middelheim hospital (Antwerp, Belgium) (Creemers and Lambrecht, 2007). We omit in this paper all practical data collection details of the case. We now and then provide numerical data to give the reader an idea of the problem dimension. In our example, the department employs six surgeons. Each of the surgeons is assigned a certain number of patients and no patient crossover between surgeons is assumed to take place. The base case deals in other words with the non-pooled capacity. Recovery occurs in an internal ward, an external ward or in the day hospital (depending on the disorder the patient is suffering from). In each of the wards 25 beds are reserved for patients of the hospital department under study. The capacity structure of the department is illustrated in Figure 18.1. Notwithstanding the fact that every patient is unique, we impose some general assumptions regarding the treatment process of a patient visiting the department. More specifically, we assume that every patient starts the treatment process with one or more consultations. Next, surgery is performed and a number of follow-up consultations is initiated. Finally the treatment process of a patient finishes and the patient leaves the hospital system. We assume that only elective surgery takes place and that the consultation process is appointment-based. Remark that it is possible to specify other patient routings (e.g. patients who refuse surgery, patients that do not longer need recovery, . . .). In this example, however, we make use of a simple patient routing structure in order to preserve the transparency of the model.

With respect to the performance measures, we are interested in the total flow time of a patient at a workstation (i.e. consultation, surgery or recovery). We define the flow time as the total waiting time plus the processing time. With respect to the waiting time of a patient, a distinction is made between the internal waiting time and the external waiting time (Vissers, Bertrand and De Vries (2001) and Hall et

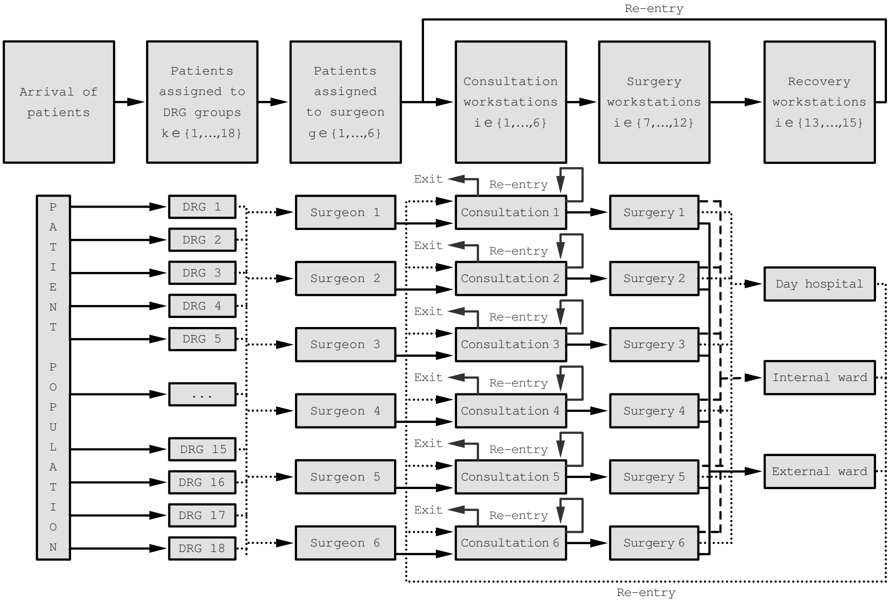


Fig. 18.1: Hospital queueing network

al. (2006)). More specifically, the internal waiting time is the time spent inside the hospital prior to receiving service (at any of the workstations). The external waiting time is the time between the making of an appointment and the arrival of a patient at the hospital. The external waiting time can also be related to the ”waiting list” phenomenon. As such, the total flow time of a patient consists of: (1) the external waiting time; (2) the internal waiting time; (3) the processing time. In the remainder of this text we will use $E[W]$ to denote the total flow time of a patient.

The data collection may be described in the following way (see also Figure 18.1). We start with a patient population (in our case we collected data on the consultation, surgery and recovery process of 3,300 patients) and divide it into groups of similar DRG’s. We construct 18 DRG groups and use index k, $k \in \{1,2,\ldots,K\}$ for further identification (refer to Roth and Van Dierdonck (1995) and van Merode, Groothuis and Hasman (2004) for a detailed treatment on patient classification methodology). Next, the patients are assigned an individual surgeon (identified using index g, $g \in \{1,2,\ldots,G\}$). Surgeons as well as recovery wards may be considered as hospital resources. We use index i, $i \in \{1,2,\ldots,I\}$ to identify these resources. The surgeons perform both consultation ($i \in \{1,2,\ldots,6\}$) as well as surgery ($i \in \{7,8,\ldots,12\}$) tasks. Recovery takes place at the day hospital ($i = 13$), the internal ward ($i = 14$) or the external ward ($i = 15$).

In what follows we develop the queueing model. First we provide the mathematical derivations required to obtain the arrival and natural process times. Next, we adapt the model to include the effects of service outages, the availability of workstations and the characteristics of the aggregate arrival process.

18.3.1 Modeling arrival rate and natural service times

The queueing model of the hospital department may be presented as a network of 12 $G/G/1$ workstations (six surgeons performing both consultation and surgery) and 3 $G/G/m$ workstations (the recovery wards). The network is an open re-entry network with stochastic routings and is modeled using the principles of the parametric decomposition approach. While other approaches are available (e.g. Brownian motion queueing models), a previous study has shown that the parametric decomposition approach works best when modeling complex hospital systems (Creemers et al., 2007).

The queue discipline adhered to at each of the stations is FCFS. Any variation in the arrival of patients (e.g. the early, late, unannounced or not showing up of patients) is presumed to be absorbed in the variance of the arrival process. The model assumes infinite buffers to exist in front of every queue. Realizing that the buffers in front of the consultation and surgery workstation correspond to their respective waiting lists, it would be incorrect to restrain them in size. In real life, if patients contact the hospital to make an appointment for a consultation or a surgery, they will be issued an appointment date no matter how far ahead in time this date might be (i.e. we assume patients not to display any balking- or reneging-behavior when arriving or abiding at the queue). Hence buffer capacities are virtually unlimited. With respect to the recovery wards, one might argue that queue capacity is in fact limited. However, there are several reasons that are able to question this assertion. Next to rendering the model highly intractable, finite buffers do not necessarily correspond to reality since shortages of bed capacity at the wards are solved at the local level and in general do not prolong the sojourn time of a patient (this of course presumes the presence of unoccupied beds somewhere in the hospital). Therefore we will assume infinite buffers at all stages of the treatment process. Considering the multiclass re-entry environment of the queueing network, aggregation of the arrival and service process is required in order to perform a decomposition-based queueing analysis.

More formally, let i ($i \in \{1,\dots,I\}$) denote the workstation in the network, let k ($k \in \{1,\dots,K\}$) denote the DRG group a patient belongs to and let g ($g \in \{1,\dots,G\}$) denote the surgeon a patient is assigned to. As such, we have KG classes of patients visiting a set of I workstations. Let the pair (k,g) denote the class of a patient (i.e. a patient of class (k,g) is assigned a surgeon g and belongs to DRG group k). Patients belonging to different classes are allowed to differ in terms of interarrival times, service times and routing. Assume interarrival times and service times of patients to be i.i.d. if they belong to one and the same class and assume them to be independently (but not necessarily identically) distributed otherwise. Let $\eta_{i(k,g)}$ denote the external arrival rate of a class (k,g) patient at workstation i (remark that external arrivals are only assumed to take place at the consultation workstations). The aggregate external arrival rate at a workstation i equals:

$$\eta_i = \sum_{k=1}^{K} \sum_{g=1}^{G} \eta_{i(k,g)}. \tag{18.1}$$

Note that expression 18.1 is a general expression, most of the time a workstation will be uniquely assigned to a single surgeon, making the summation over g redundant.

We assume that the interarrival times of the external arrivals are exponentially distributed. Such an assumption poses only a slight restriction on the accuracy of the model while it has been shown by Palm (1943) and Khinchin (1960) that the sum of a large numbers of independent renewal processes (i.e. the arrival processes of the different classes of patients) will tend to a Poisson process. Considering the multitude of classes of patients, the approximation of the aggregate external arrival process by means of a Poisson process should be accurate. In addition, Lariviere and Van Mieghem (2004) showed that the assumption of exponential interarrival times is reasonable in many service systems.

Let $\gamma_{i(k,g)}$ denote the expected number of visits a class (k,g) patient will make to workstation i (remark that only the consultation workstations are assumed to be visited more than once). The aggregate arrival rate of patients at the consultation level equals:

$$\lambda_i = \sum_{k=1}^{K} \sum_{g=1}^{G} \eta_{i(k,g)} \gamma_{i(k,g)}, \; \forall i \in \{1,2,\dots,6\}. \tag{18.2}$$

Note that in contrast to the aggregate external arrival rate, which was assumed to be Poisson-distributed, the aggregate arrival rate (at each of the workstations) is allowed to follow a general distribution. Further define the routing matrix R in which the elements r_{ij} indicate the probability of a patient to travel from station i to station j after service completion at station i. Adhering to standard conventions, we establish a node (of index $i = 0$) from which external arrivals originate and which also serves as a sink for patients leaving the hospital system. Let r_{i0} indicate the probability of leaving the system when departing from station i. Conversely r_{0i} implies the probability of an external arrival occurring at station i. The probabilities r_{ij} can be expressed as the the proportion of the arrivals at station i that travel towards station j. When assuming the stability of the queueing network, the law of conservation of flows (what comes in, must go out) dictates:

$$r_{i0} = r_{0i} = \frac{\eta_i}{\lambda_i}\ \forall i \in \{1,2,\ldots,6\}. \tag{18.3}$$

With respect to the surgery workstations, each patient visiting the hospital department is subjected to surgery exactly once. As such, one can infer that:

$$\lambda_i = \eta_i,\ \forall i \in \{7,8,\ldots,12\}. \tag{18.4}$$

Hence the probability of transition from the consultation to the surgery level may be defined as:

$$r_{ij} = \frac{\eta_i}{\lambda_i},\ \forall i \in \{1,2,\ldots,6\},\ j = i+6. \tag{18.5}$$

Finally, at the consultation level, the probability of re-entry equals:

$$r_{ii} = 1 - (r_{i0} + r_{ij}) = 1 - \frac{2\eta_i}{\lambda_i},\ \forall i \in \{1,2,\ldots,6\},\ j = i+6. \tag{18.6}$$

The routing probabilities of transferring from a surgery workstation $i,\ i \in \{7,8,\ldots,12\}$ towards a recovery ward $j,\ j \in \{13,14,15\}$ is obtained as follows:

$$r_{ij} = \frac{\lambda_j^{(i)}}{\lambda_i},\ \forall i \in \{7,8,\ldots,12\},\ \forall j \in \{13,14,15\}, \tag{18.7}$$

where $\lambda_j^{(i)}$ is the empirically observed arrival rate of patients at recovery workstation $j,\ j \in \{13,14,15\}$ originating from surgery workstation $i,\ i \in \{7,8,\ldots,12\}$. As such, the arrival rates at recovery equal:

$$\lambda_j = \sum_{i=7}^{12} \lambda_j^{(i)},\ \forall j \in \{13,14,15\}. \tag{18.8}$$

From this we obtain:

$$r_{ij} = \frac{\lambda_i^{(j+6)}}{\lambda_i},\ \forall i \in \{13,14,15\},\ \forall j \in \{1,2,\ldots,6\}. \tag{18.9}$$

All other routing probabilities stem directly from the structure of the model. A schematic summary of the routing matrix R is presented in Table 18.1. Note that

Table 18.1: Schematic summary of the routing matrix R

i / j	0	1-6	7-12	13-15
0	0	$\frac{\eta_j}{\lambda_j}$	0	0
1-6	$\frac{\eta_i}{\lambda_i}$	$\delta_{ij}\left(1-\frac{2\eta_i}{\lambda_i}\right)$	$\delta_{ij}\left(\frac{\eta_i}{\lambda_i}\right)$	0
7-12	0	0	0	$\frac{\lambda_j^{(i)}}{\lambda_i}$
13-15	0	$\frac{\lambda_i^{(j+6)}}{\lambda_i}$	0	0

$(\delta_{ij}=1)$ if at least one of the patient classes travels from station i to station j and $(\delta_{ij}=0)$ otherwise.

Remark that other routing structures give rise to other routing probabilities. The routing structure and corresponding equations discussed in this section are only valid under the previously imposed assumptions concerning patient flow.

With respect to the service times, let $f_{i(k,g)}(x)$ denote the natural service time probability density function of a class (k,g) patient visiting workstation i. Have $\frac{1}{\nu_{i(k,g)}}$ and $\sigma^2_{\nu_{i(k,g)}}$ represent the average natural service time for a class (k,g) patient at workstation i and its variance respectively. The natural process time excludes random interruptions, absences and any other external influence. Assume service times of different classes to be independent but not necessarily identically distributed. The probability that a randomly picked unit in front of the workstation is of class (k,g) is given by $\frac{\lambda_{i(k,g)}}{\lambda_i}$, where $\lambda_{i(k,g)}$ is the total arrival rate of class (k,g) patients at workstation i. Define the probability function of the aggregate natural service times at station i as follows:

$$f_i(x)=\sum_{k=1}^{K}\sum_{g=1}^{G}\frac{\lambda_{i(k,g)}}{\lambda_i}f_{i(k,g)}(x). \tag{18.10}$$

As a result the average natural service time requirement of a unit in front of the workstation amounts to:

$$\frac{1}{\nu_i}=\sum_{k=1}^{K}\sum_{g=1}^{G}\frac{\lambda_{i(k,g)}}{\lambda_i}\frac{1}{\nu_{i(k,g)}}. \tag{18.11}$$

When observing the variance of the aggregate natural service process, one can deduce that:

$$\begin{aligned}\sigma^2_{\nu_i}&=\sum_{k=1}^{K}\sum_{g=1}^{G}\frac{\lambda_{i(k,g)}}{\lambda_i}\int\left(x-\frac{1}{\nu_i}\right)^2 f_{i(k,g)}(x)\,dx,\\&=-\frac{1}{\nu_i^2}+\sum_{k=1}^{K}\sum_{g=1}^{G}\frac{\lambda_{i(k,g)}}{\lambda_i}\left(\sigma^2_{\nu_{i(k,g)}}+\frac{1}{\nu^2_{i(k,g)}}\right).\end{aligned} \tag{18.12}$$

We refer to $\sigma_{v_i}^2$ as a measure of the natural variability of the aggregate process times at workstation i. The same result was obtained by Whitt (1983) and has widely been adopted in literature (Whitt (1999b) and Haskose, Kingsman and Worthington (2002)).

18.3.2 Variability from preemptive and nonpreemptive outages

With respect to service outages in healthcare, a large body of literature exists. Outages in a hospital setting have been the subject of discussion in Babes and Sarma (1991), Liu and Liu (1998a), Chisholm, Collison, Nelson and Cordell (2000) and Chisholm, Dornfeld, Nelson and Cordell (2001) among others. There is a consensus on the harmful effects of outages on patient flow times as well as on the quality of service. Outages result in congestion, unstable schedules and most importantly in overtime for staff members. We refer to Easton and Goodale (2005) for an excellent treatment of this issue. In this section, we focus on unplanned absences of medical staff and interruptions during service operations. Unplanned absences and interruptions during service activities have a major impact on flow times. Doctors and medical staff face various obligations which they have to attend to (making morning rounds, answering phones, patient check-ups, daily management, . . .). In addition doctors often combine a hospital job and private consultation. These phenomena may cause a variable arrival pattern at the hospital (Liu et al., 1998a) and may lead to interruptions during the treatment process (Chisholm et al. (2000 and 2001) and Easton et al. (2005)). It is clear that hospital environments are characterized by substantial amounts of variability. As is argued in the literature (Hopp et al., 2000), variability induces waiting times. While in service industries variability cannot be countered by means of inventory in the traditional sense, patients will have to wait until capacity becomes available (Vissers et al. (2001), Vandaele and De Boeck (2003a) and Sethuraman and Tirupati (2005)). Besides the time buffer, hospitals often have to rely on a capacity buffer to mitigate the impact of variability and to maintain required service levels. In order to model service processes liable to outages, queueing theory proves to be an ideal tool. With respect to service outages and server unreliability, we face a vast amount of queueing literature. Surveys on the machine interference problem and server unreliability may be found in Stecke and Aronson (1985) and Haque and Armstrong (2007). Unreliable servers are often modeled using vacation models. Over the past decades, queueing systems with server vacations have received a lot of attention in the queueing literature. Vacation models observe the queueing behavior of systems in which the server begins a vacation (i.e. becomes unavailable) when certain conditions are met. For instance, imagine a doctor's office that has opening hours on Tuesday afternoons and on Friday evenings. On Tuesday, after service completion of the last patient, the doctor leaves on a "vacation" until Friday evening at which time service is resumed. At the end of service on Friday, a vacation is initiated until next Tuesday afternoon. We illustrate this process in Figure 18.2. Next to the modeling of planned absences (e.g.

a working schedule), vacation models may also be used to model unplanned server interruptions (e.g. a doctor who is called away for an emergency). A wide variety of vacation models exists. For a general overview, we refer to Doshi (1986), Takagi (1988) and Tian and Zhang (2006). In this work, however, we do not focus on vacation models. Instead, we consider an alternative, more intuitive approach to model service outages. This approach was first suggested by Hopp and Spearman (2000). In their work, Hopp and Spearman propose a transformation of the service process times to account for service outages. The results of Hopp and Spearman are widely accepted in the literature. In this work, we develop new expressions to model the impact of service outages that are peculiar to healthcare systems. In what follows, we first discuss the difference between preemptive and nonpreemptive outages. Next, we provide the means to model them.

18.3.2.1 Outages, classification and impact

As was indicated previously, the service process of a patient may be interrupted or postponed. These outages will increase the natural service times. We call these increased, adjusted service times, effective processing times. It is the total time “seen” or “experienced” by a patient at a workstation. The effective process time random variable is of primary interest to determine flow times.

We distinguish between preemptive and nonpreemptive outages. Preemptive and nonpreemptive outages will impact the service process and will give rise to increased levels of traffic intensity (resulting in the so-called effective utilization rate or effective traffic intensity).

Let us first discuss the nonpreemptive outages. Nonpreemptive outages typically occur between jobs, rather than during jobs. They occur at the beginning of each service session (i.e. at the start of a consultation work shift) whenever a doctor or another member of the medical staff is absent (e.g. due to late arrival). We may refer to such an outage as unplanned absences and define the mean and variance of the amount of time absent as $\frac{1}{\mu_s}$ and σ_s^2 respectively (i.e. absence times are allowed to follow a general distribution). Furthermore we assume an average number of patients (represented by n) to arrive in between two consecutive absences. This is an important feature of the model. Indeed, n may be considered as the number of patients in a service session (e.g. a consultation work shift). Each start of a service

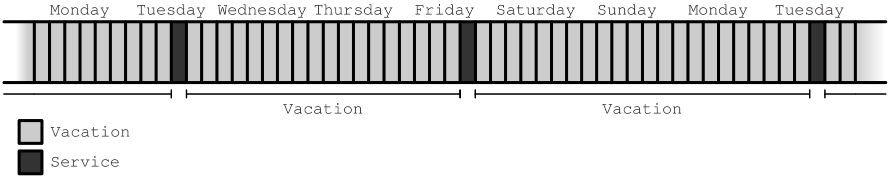


Fig. 18.2: Illustration of a vacation model

session may induce a delay due to an absence. In other words, the number of patients in a service session is a decision variable and is comparable to a lot sizing decision. Evaluating different service session sizes (i.e. different values of n) may provide key managerial insights. We will address this issue in an upcoming section.

Next to nonpreemptive outages, we also allow for preemptive outages to take place. Preemptive outages occur whenever a doctor is interrupted during a consultation activity. These interruptions will be modeled in an approach which builds on the tradition set by Hopp and Spearman (2000). They are characterized by a Mean Time To Interrupt (τ_f) and a Mean Time To Resolve (τ_r). The model presented in Hopp and Spearman (2000) presumes interrupts to occur only during actual service time. However, in a hospital setting it is not inconceivable that interrupts take place during the resolve time induced by a previous interrupt as well. For instance, if the service process of a patient is interrupted by a phone call, it is still possible for a doctor to be called away for an emergency, to receive another call, ...

In what follows, we present the main results on nonpreemptive as well as preemptive outages. In a final subsection, we present results on the joint occurrence of nonpreemptive and preemptive outages. In order to maintain transparency of the model and of notation, we impose the following assumptions: (1) service outages only occur at the consultation level (i.e. only workstations i, $i \in \{1,2,\ldots,6\}$ are affected); (2) for each of the surgeons, the impact of outages is identical (i.e. $\frac{1}{\mu_s}$, σ_s^2, n, τ_f and τ_r remain the same for each of the workstations at the consultation level).

18.3.2.2 Nonpreemptive outages

We define a nonpreemptive outage to occur whenever the succession of two events is based on the number of services performed in between (hence, setups, rework, maintenance, ... are all extensions that are able to capitalize on the technique discussed in this section). Applied to our setting, we have that n patients are treated (on average) in between two consecutive absence possibilities. Assume that the length of services and absence times does not depend on the service history (i.e. they are independent of prior services and absence times). The absence times themselves are distributed following a probability density function $f_s(x)$. The average absence time and its variance are represented by $\frac{1}{\mu_s}$ and σ_s^2. The service time of the n^{th} patient includes part service time, part absent time. We refer to the service time of the n^{th} patient as the combined service time. We illustrate these concepts in Figure 18.3.

One can consider the services that are preceded by an absent period as a separate class of patients that has a probability $\frac{1}{n}$ of randomly being picked in front of the workstation. The other services as a whole have a probability $\left(\frac{(n-1)}{n}\right)$ of randomly being picked. Therefore, we can define the mean aggregate service times including the effect of absence times as follows:

$$\begin{aligned}\frac{1}{\upsilon_i} &= \left[\left(\tfrac{n-1}{n}\right)\sum_{k=1}^{K}\sum_{g=1}^{G}\tfrac{\lambda_{i(k,g)}}{\lambda_i}\int f_{i(k,g)}(x)\,x\,dx\right]+\\ &\quad\left[\tfrac{1}{n}\sum_{k=1}^{K}\sum_{g=1}^{G}\tfrac{\lambda_{i(k,g)}}{\lambda_i}\iint f_{i(k,g)}(x)\,f_s(y)\,(x+y)\,dydx\right],\\ &= \frac{1}{\nu_i}+\frac{1}{n\mu_s}.\end{aligned} \tag{18.13}$$

With respect to the variance of the aggregate service time (including absence times) at the consultation workstations we develop the following expression:

$$\begin{aligned}\sigma^2_{\upsilon_i} &= \left[\left(\tfrac{n-1}{n}\right)\sum_{k=1}^{K}\sum_{g=1}^{G}\tfrac{\lambda_{i(k,g)}}{\lambda_i}\int f_{i(k,g)}(x)\left(x-\tfrac{1}{\upsilon_i}\right)^2 dx\right]+\\ &\quad\left[\tfrac{1}{n}\sum_{k=1}^{K}\sum_{g=1}^{G}\tfrac{\lambda_{i(k,g)}}{\lambda_i}\iint f_{i(k,g)}(x)\,f_s(y)\left(x+y-\tfrac{1}{\upsilon_i}\right)^2 dydx\right],\\ &= \sigma^2_{\nu_i}+\tfrac{\sigma_s^2}{n}+\tfrac{1}{\mu_s^2}\left(\tfrac{n-1}{n^2}\right).\end{aligned} \tag{18.14}$$

The above expression is equivalent to that of Hopp and Spearman (2000) and is valid under the assumption that the combined service times as well as ordinary service times are independently distributed.

18.3.2.3 Preemptive outages

We refer to service interruptions as preemptive outages. Doctors being called away on emergencies, answering phone calls, ... are typical examples. The average time between two consecutive interrupts is defined as τ_f whereas τ_r refers to the average time it takes to resolve an interruption. Preemptive outages prove to be more difficult to model while they occur after the elapsing of a variable amount of time (i.e. a mean time to interrupt τ_f), rather than after a number of patients being processed. Under the assumption that the time between two consecutive interrupts is exponentially distributed, expressions for mean and variance have been obtained. With respect to preemptive outages, we make a distinction between two different scenarios. On the one hand, one might presume preemptive outages to occur only during

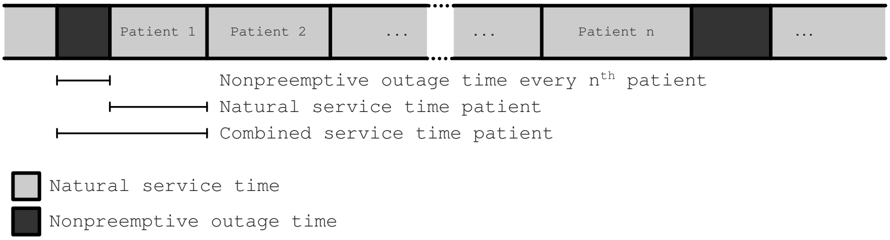

Patient 1
Patient 2
...
...
Patient n
...
Nonpreemptive outage time every n^th patient
Natural service time patient
Combined service time patient
Natural service time
Nonpreemptive outage time

Fig. 18.3: The combined service time

actual service time. As such preemptive outages do not take place during the resolve times induced by previous outages. Remark that this does not imply that the service process of a single patient cannot be interrupted more than once. On the other hand, one might assume preemptive outages to occur during resolve times as well (e.g. as indicated previously, doctors may be be interrupted when already engaged in resolving a previous interrupt). While this latter instance can be seen as an extension of the former, we will first discuss outages occurring exclusively during actual service time. Define $\tau_{r_{0_j}}$ as the resolve time of the j^{th} preemptive outage that occurred during the service process of one and the same patient. The mean and variance of the resolve times are given by τ_r and σ_r^2. In addition, resolve times of different outages are assumed to be i.i.d.. The service process of a patient thus faces the probability of encompassing several interrupts that prolong its service duration. The service time of a patient (including interrupts) at a workstation i can be expressed as:

$$\frac{1}{\omega_i} = \frac{1}{\nu_i} + \sum_{j=1}^{J_0} \tau_{r_{0_j}}. \tag{18.15}$$

As such, the random variable $\frac{1}{\omega_i}$ incorporates both the natural service time $\frac{1}{\nu_i}$ as well as the resolve times of interrupts that occurred during service. Moreover, J_0 denotes the number of preemptive outages that occurred during the service process of a unit. J_0 is a random variable that follows a Poisson distribution (i.e. we assume the time between two consecutive interrupts to be exponentially distributed) and its mean and variance both equal $\left(\frac{1}{(\nu_i \tau_f)}\right)$. We face a sum of random variables (the resolve times $\tau_{r_{0_j}}$) in which the number of random variables (the number of interrupts J_0), is a random variable itself. Assume that J_0 and $\tau_{r_{0_j}}$ $(\forall j \in \mathbb{N})$ are i.i.d. variables. In addition assume the mean as well as the variance of $\tau_{r_{0_j}}$ to be equal for all $j \in \mathbb{N}$. Therefore, the mean and variance of the sum of J_{i_0} random variables $\tau_{r_{0_j}}$ can be expressed as (Dudewicz and Mishra, 1988):

$$E\left[S_0\right] = E\left[J_0\right] E\left[\tau_{r_{0_j}}\right], \tag{18.16}$$

$$\sigma_{S_0}^2 = E\left[J_0\right] \sigma_r^2 + E\left[\tau_{r_{0_j}}\right]^2 \sigma_{J_0}^2, \tag{18.17}$$

where S_0 is the random variable representing the sum of J_0 resolve times $\tau_{r_{0_j}}$. In other words we have that:

$$S_0 = \sum_{j=1}^{J_0} \tau_{r_{0_j}}. \tag{18.18}$$

The mean and variance of the sum of resolve times can be defined as:

$$E\left[S_0\right] = \frac{1}{\nu_i} \frac{\tau_r}{\tau_f}, \tag{18.19}$$

$$\sigma_{S_0}^2 = \frac{1}{\nu_i}\frac{\sigma_r^2 + \tau_r^2}{\tau_f}. \tag{18.20}$$

The mean aggregate service time including the effect of interrupts may be expressed as:

$$E\left[\frac{1}{\omega_i}\right] = \frac{1}{\nu_i}\frac{\tau_f + \tau_r}{\tau_f}. \tag{18.21}$$

This corresponds to the expression presented in Hopp and Spearman (2000) in which the natural service time is divided by an availability factor in order to incorporate the effect of interrupts. Next we have a look at the variance of the service times including the effect of preemptive outages during service time. We start with the approximation of the second moment:

$$E\left[\left(\frac{1}{\omega_i}\right)^2\right] = \left(\sigma_{\nu_i}^2 + \frac{1}{\nu_i^2}\right)\left(1 + \frac{\tau_r}{\tau_f}\right)^2 + \sigma_{S_0}^2. \tag{18.22}$$

Using the expression for the second moment we obtain the variance of the service times including the effect of interrupts:

$$\sigma_{\omega_i}^2 = \sigma_{\nu_i}^2\left(1 + \frac{\tau_r}{\tau_f}\right)^2 + \sigma_{S_0}^2. \tag{18.23}$$

This expression once more matches the formula derived in Hopp and Spearman (2000). The above expressions hold if and only if the Poisson-distributed preemptive outages take place during service itself. In what follows, we relax this assumption and allow for interrupts to take place during the resolve times induced by previous interrupts.

In order to approach this problem, we divide the interrupts into different sets. Let l ($l \in \mathbb{N}$) denote the set index. We define $\tau_{r_{l_j}}$ to be the resolve time of the j^{th} interrupt belonging to the set of index l (i.e. the interrupt is said to be of order l). Without loss of generality assume that interrupts of order 0 occurred during actual service, interrupts of order 1 occurred during the resolve times of interrupts of order 0, ... In general, interrupts of order l took place during the resolving of interrupts of order $(l-1)$. Figure 18.4 provides further insight. In addition define S_l as the sum of resolve times corresponding to interrupts of order l. We have that:

$$S_l = \sum_{j=0}^{J_l} \tau_{r_{l_j}}, \tag{18.24}$$

where J_l is the number of interrupts belonging to the set of index l. J_l follows a Poisson distribution and its mean and variance equal:

$$E\left[J_l\right] = \sigma_{J_l}^2 = \frac{1}{\nu_i \tau_f}\left(\frac{\tau_r}{\tau_f}\right)^l. \tag{18.25}$$

One can infer that:

$$E\left[S_l\right] = \frac{\tau_r}{\nu_i \tau_f}\left(\frac{\tau_r}{\tau_f}\right)^l, \tag{18.26}$$

$$\sigma^2_{S_l} = \frac{1}{\nu_i \tau_f}\left(\frac{\tau_r}{\tau_f}\right)^l \left(\sigma_r^2 + \tau_r^2\right). \tag{18.27}$$

Using the same reasoning applied previously, one can express the mean aggregate service time including the effect of all order interrupts as follows:

$$E\left[\frac{1}{\omega_i}\right] = \frac{1}{\nu_i}\frac{\tau_f}{\tau_f - \tau_r}. \tag{18.28}$$

Using these parameters, the second moment is expressed as:

$$E\left[\left(\frac{1}{\omega_i}\right)^2\right] = \left(\sigma^2_{\nu_i} + \frac{1}{\nu_i^2}\right)\left[1 + 2\frac{\tau_r}{\tau_f - \tau_r} + \left(\frac{\tau_r}{\tau_f - \tau_r}\right)^2\right] + \frac{1}{\nu_i}\frac{\sigma_r^2 + \tau_r^2}{\tau_f - \tau_r}. \tag{18.29}$$

As a result, the variance of the service time at a workstation i (including the impact of all order interrupts) is given by:

$$\sigma^2_{\omega_i} = \frac{\tau_f^2 \sigma^2_{\nu_i} + \frac{1}{\nu_i}\left(\tau_f - \tau_r\right)\left(\sigma_r^2 + \tau_r^2\right)}{\left(\tau_f - \tau_r\right)^2}. \tag{18.30}$$

18.3.2.4 Combining preemptive and nonpreemptive outages

In many hospital settings, both preemptive and nonpreemptive outages may surface. While it is impossible to interrupt the service process in the instance of a nonpreemptive outage (e.g. a doctor who arrives late), we only consider the case in which both types of outages cannot occur simultaneously. The average service time incorporating this combined effect at a workstation i can be expressed as:

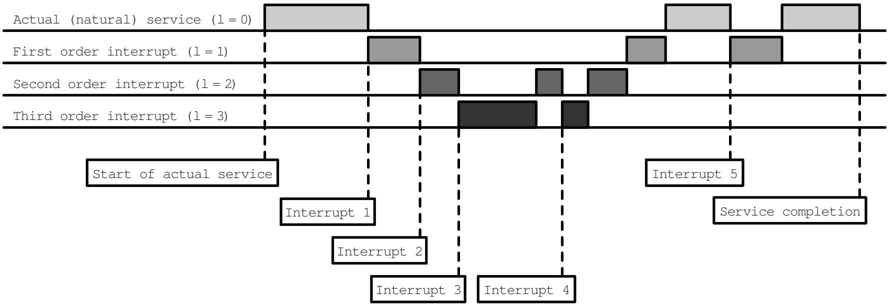

Actual (natural) service (l = 0)
First order interrupt (l = 1)
Second order interrupt (l = 2)
Third order interrupt (l = 3)
Start of actual service
Interrupt 1
Interrupt 2
Interrupt 3
Interrupt 4
Interrupt 5
Service completion

Fig. 18.4: Interrupted service process of a single patient

$$\begin{aligned}\tfrac{1}{\psi_i} &= \left[\left(\tfrac{n-1}{n}\right)\sum_{k=1}^{K}\sum_{g=1}^{G}\tfrac{\lambda_{i(k,g)}}{\lambda_i}\int f_{i_{f(k,g)}}(x)\,x\,dx\right]+ \\ &\quad \left[\tfrac{1}{n}\sum_{k=1}^{K}\sum_{g=1}^{G}\tfrac{\lambda_{i(k,g)}}{\lambda_i}\iint f_{i_{f(k,g)}}(x)\,f_s(y)\,(x+y)\,dydx\right], \\ &= E\left[\tfrac{1}{\omega_i}\right]+\tfrac{1}{n\mu_s},\end{aligned} \tag{18.31}$$

where $f_{i_{f(k,g)}}(x)$ is the probability density function of consultation service times of a class (k,g) patient at a workstation i including the effect of all order interrupts. Its mean and variance are given by $E\left[\frac{1}{\omega_i}\right]$ and $\sigma^2_{\omega_i}$ respectively. We refer to $\frac{1}{\psi_i}$ as the effective service time while it equals the service time experienced by the patient (and as such includes the impact of outages). The variance of the effective service times at a workstation i may be expressed as:

$$\begin{aligned}\sigma^2_{\psi_i} &= \left[\left(\tfrac{n-1}{n}\right)\sum_{k=1}^{K}\sum_{g=1}^{G}\tfrac{\lambda_{i(k,g)}}{\lambda_i}\int f_{i_{f(k,g)}}(x)\left(x-\tfrac{1}{\psi_i}\right)^2 dx\right]+ \\ &\quad \left[\tfrac{1}{n}\sum_{k=1}^{K}\sum_{g=1}^{G}\tfrac{\lambda_{i(k,g)}}{\lambda_i}\iint f_{i_{f(k,g)}}(x)\,f_s(y)\left(x+y-\tfrac{1}{\psi_i}\right)^2 dydx\right], \\ &= \sigma^2_{\omega_i}+\tfrac{\sigma_s^2}{n}+\tfrac{1}{\mu_s^2}\left(\tfrac{n-1}{n^2}\right).\end{aligned} \tag{18.32}$$

These results allow us to take service outages into account when assessing hospital performance measures.

18.3.2.5 Including the time availability of workstations

It is well known that many services do not operate continuously over time. Consultation and surgery typically operate during certain time intervals (service sessions) which means that only a proportion of the total available time can be used effectively. Vacation models are often applied to solve this problem. Another way to handle the problem is to rescale all service processing times so that they fit a preset uniform time scale. In this study we agreed on a 24 hours per day, 7 days per week time scale (basically because this is the appropriate time scale for recovery processes). Let A_i denote the availability of workstation i; A_i represents the available time in proportion to the preset uniform time scale. For instance, if a workstation operates only 6 hours per day, then the availability equals 25%.

When rescaling the service times established in the previous sections, we obtain the total effective service times:

$$\frac{1}{\mu_i} = \frac{1}{A_i \psi_i}, \forall i \in \{1,2,\ldots,6\}, \tag{18.33}$$

$$\frac{1}{\mu_i} = \frac{1}{A_i \nu_i} \forall i \in \{7,8,\ldots,15\}, \tag{18.34}$$

$$\sigma_i^2 = \frac{\sigma_{\psi_i}^2}{A_i^2}, \forall i \in \{1,2,\ldots,6\}, \tag{18.35}$$

$$\sigma_i^2 = \frac{\sigma_{\nu_i}^2}{A_i^2} \forall i \in \{7,8,\ldots,15\}. \tag{18.36}$$

The above procedure results in the total effective service times including natural process time, the effect of outages and the impact of availability of workstations. The mean total effective service time and its variance can now be used to compute the squared coefficient of variation of the service times:

$$C_{s_i}^2 = \sigma_i^2 \mu_i^2. \tag{18.37}$$

18.3.2.6 Squared coefficient of variation of the aggregate arrival process

In order to approximate the parameters of the aggregate arrival process, some more challenging arithmetics are needed. It was pointed out by Albin (1984) that if at least one of the interarrival time distributions, constituting the arrival process, does not stem from a Poisson process, the resulting aggregate interarrival times do no longer hold the property of independence. As a result the analytical analysis of the aggregate arrival process becomes highly intractable. Therefore approximations will be adopted to assess the variance and, more important, the squared coefficient of variation of the aggregate arrival process. The squared coefficients of variation of the aggregate arrivals at the different workstations will be extracted using a technique which was pioneered by Shanthikumar and Buzacott (1981). This technique implies the use of a set of linear equations which has to be solved in order to obtain the squared coefficients of variation of the arrivals. This approach is widely adopted in literature (Askin, 1993) and was later generalized by Lambrecht et al. (1998). Using the technique that was outlined in Lambrecht et al. (1998), we are given a set of I equations:

$$-\sum_{i=1}^{I} \lambda_i r_{ij}^2 (1-\rho_i^2) C_{a_i}^2 + \lambda_j C_{a_j}^2 = \sum_{i=1}^{I} \lambda_i r_{ij} (r_{ij} \rho_i^2 C_{s_i}^2 + 1 - r_{ij}) + \eta_j C_{a_{\eta_j}}^2, \tag{18.38}$$

where η_j and $C_{a_{\eta_j}}^2$ denote the rate and squared coefficient of variation of the aggregate external arrival process at station j respectively. In addition, ρ_i represents the effective traffic intensity at workstation i and equals $\frac{\lambda_i}{\mu_i}$. While all elements except the I squared coefficients of variation are known, we are presented with a system of I equations yielding I unknowns. Solving this set of linear equations provides us with the I unknown squared coefficients of variation (i.e. $C_{a_i}^2; \forall i \in \{1,\ldots,I\}$).

With all model parameters firmly defined, we now have a solid base to carry out the performance evaluation of the hospital department. In the upcoming section we discuss a numerical example of the model presented above and provide some practical applications.

18.4 Applications

In this section, we discuss a numerical example using the queueing model described in the previous section. Next, we illustrate the devastating impact of service interruptions on patient flow times. Subsequently, we show the potential gains obtained by pooling hospital resources. Finally, we present an optimization model to determine the optimal number of patients to be treated during a service session.

18.4.1 Numerical example

The numerical example presented in this section builds on data gathered at the orthopedic department of the Middelheim hospital in Antwerpen (Belgium). Using these empirical data as inputs, the flow time of patients at the hospital department may be assessed using so-called flow time expressions. A variety of flow time expressions are available in the queueing literature. A previous study has shown the Kingman equation to yield accurate results when assessing the flow times of patients in complex hospital systems (Creemers et al., 2007). As such, in the remainder of this article, we will use the Kingman equation to determine patient flow times. With respect to the Kingman equation, one can define the expected flow time of a patient at workstation i as follows (Hopp et al., 2000):

$$E\left[W_i\right] = \left(\frac{C_{a_i}^2 + C_{s_i}^2}{2}\right)\left(\frac{\rho_i^{\sqrt{2(m_i+1)}-1}}{m_i\left(1-\rho_i\right)}\right)\frac{1}{\mu_i} + \frac{1}{\mu_i}, \tag{18.1}$$

where m_i denotes the number of parallel servers at workstation i ($m_i = 25\ \forall i \in \{13,14,15\}$). If only a single server is present (i.e. at workstations i, $i \in \{1,2,\ldots,12\}$), no pooling is assumed to take place and the formula reduces to (Kingman, 1962):

$$E\left[W_i\right] = \left(\frac{C_{a_i}^2 + C_{s_i}^2}{2}\right)\left(\frac{\rho_i}{1-\rho_i}\right)\frac{1}{\mu_i} + \frac{1}{\mu_i}. \tag{18.2}$$

Using the empirical data, resulting flow times at each of the workstations are obtained. The results are presented in Table 18.2 and Table 18.3 (all results are expressed in minutes unless indicated otherwise). While no waiting occurs at the wards (i.e. the process of recovery takes place immediately after surgery) the performance measures of workstations 13 to 15 are not included here.

Table 18.2: Summary Table of the model results (workstations 1 to 6)

i	1	2	3	4	5	6
$\frac{1}{\psi_i}$	24.85	24.85	24.85	24.85	24.85	24.85
$\frac{1}{\mu_i}$	310.7	690.4	310.7	167.9	155.3	248.5
$C^2_{s_i}$	1.334	1.334	1.334	1.334	1.334	1.334
$\frac{1}{\lambda_i}$	329.8	741.5	317.0	174.5	167.5	268.8
$C^2_{a_i}$	1.026	1.418	1.051	0.759	0.752	0.952
A_i	0.080	0.036	0.080	0.148	0.160	0.100
ρ_i	0.942	0.931	0.980	0.962	0.927	0.925
$E[W_i]$ (days)	4.360	9.402	12.90	3.219	1.547	2.593

Table 18.3: Summary Table of the model results (workstations 7 to 12)

i	7	8	9	10	11	12
$\frac{1}{\nu_i}$	110.0	96.20	89.17	57.50	56.35	93.18
$\frac{1}{\mu_i}$	1048	2004	1351	845.7	593.2	1035
$C^2_{s_i}$	0.266	0.406	0.203	0.171	0.165	0.274
$\frac{1}{\lambda_i}$	1,111	2,111	1,380	883.4	620.5	1,073
$C^2_{a_i}$	1.089	1.121	1.074	1.058	1.068	1.070
A_i	0.105	0.048	0.066	0.068	0.095	0.090
ρ_i	0.943	0.950	0.979	0.957	0.956	0.965
$E[W_i]$ (days)	8.907	21.38	29.42	8.674	5.918	14.14

With respect to consultation, no distinction was made between the different surgeons. One can observe that the effective service time (including the effect of interrupts and absences) amounts to 24.85 minutes (the natural service time amounting to 15 minutes). The coefficient of variation equals 1.334 (the natural coefficient of variation amounting to 0.6386). Arrival rates and their variances depend on the number of patients visiting each surgeon. The utilization rates of the surgeons are all very high, which translates into significant patient flow times varying from 1.5 days to 12.9 days.

Similar observations may be made with respect to surgery. Here we allow surgeons to have different processing times depending on the type of surgery they perform. In addition, observe the significantly longer waiting times for patients at the surgery level.

18.4.2 The impact of interrupts

The impact of interrupts on medical practice has been observed by Harvey, Jarrett and Peltekian (1994), Lehaney, Clarke and Paul (1999), Chisholm et al. (2001), France, Levin, Hemphill, Chen, Rickard, Makowski, Jones and Aronsky (2005), Volpp and Grande (2006), Tucker and Spear (2006) and Gabow, Karkhanis, Knight, Dixon, Eiser and Albert (2006) among others. All agree on the detrimental effects of interrupts on patient flow time. In order to demonstrate these detrimental effects, we present a number of scenarios in which we gradually reduce the impact of interrupts. We build on the setting of the hospital department discussed previously. To maintain transparency, we focus on a single consultation workstation (i.e. the only workstations that are susceptible to interrupts during the service process). We adjust the mean time to interrupt (i.e. τ_f) at this workstation to assess the varying impact of interrupts (all other model parameters remain unchanged). The results are given in Table 18.4. Note that we used the third workstation to study the impact of various degrees of interrupts (the results corresponding to the numerical example are indicated in bold). Figure 18.5 illustrates the phenomenon graphically.

Table 18.4: Impact of interrupts (expressed in minutes) on patient flow time (expressed in days) at a single workstation

τ_f	$E[W]$	ρ	τ_f	$E[W]$	ρ	τ_f	$E[W]$	ρ
10.4	183.2	0.998	11.6	16.24	0.984	18	4.433	0.943
10.5	93.58	0.997	11.8	14.35	0.982	20	3.393	0.936
10.6	63.28	0.995	**12.0**	**12.90**	**0.980**	25	3.288	0.924
10.7	48.05	0.994	12.5	10.43	0.975	30	2.968	0.916
10.8	38.88	0.993	13.0	8.880	0.971	40	2.652	0.907
10.9	32.76	0.992	14.0	7.029	0.963	60	2.401	0.897
11.0	28.38	0.990	15.0	5.966	0.957	80	2.294	0.893
11.2	22.54	0.988	16.0	5.276	0.952			
11.4	18.82	0.986	17.0	4.791	0.947			

It is clear that heavy traffic systems (i.e. systems which operate under high workload) benefit greatly from even a small reduction in utilization rate. Unfortunately, only limited means are available to achieve such a reduction in utilization rate. A variety of options arise:

- The most obvious way to reduce the effective utilization is process improvement. Continuous improvement and six sigma programs are very beneficial. Reducing the frequency of interrupts can be classified in this category.
- Expand capacity; hospital resources such as operating theatres, scanners and other equipment are often operating at maximum capacity. Expanding capacity would be an effective means to reduce hospital workload. However, expanding

capacity is often very expensive or is simply impossible (e.g. due to legal constraints).
- Limit patient volumes; a reduction in hospital workload might also be achieved by limiting the amount of patients receiving treatment. Pursuing this option however, results in loss of hospital income and a reduced level of service.

In the literature, valuable insights are provided that offer guidance in the quest to reduce the impact of interrupts. For instance, Harvey et al. (1994) suggest the pooling of paging of doctors (next to telephone calls, paging calls are one of the largest sources of interrupts) in order to decrease variability in individual paging patterns. France et al. (2005) propose the use of information systems (e.g. an electronic whiteboard) and team training to enhance performance. Tucker et al. (2006) suggest the redesign of treatment processes (e.g. outsourcing of administrative tasks) in order to make service more robust against preemptive outages. In addition Tucker et al. (2006) and Volpp et al. (2006) propose the filtering of non-urgent communication towards medical staff. These and other practical guidelines enable hospital decision makers to minimize the impact of interrupts on the service process.

18.4.3 The impact of pooling

Pooling refers to the aggregation (consolidation) of the demand from multiple items into one, such that the consolidated demand can be satisfied from a single buffer. More specifically, capacity pooling refers to the idea of sharing available capacity

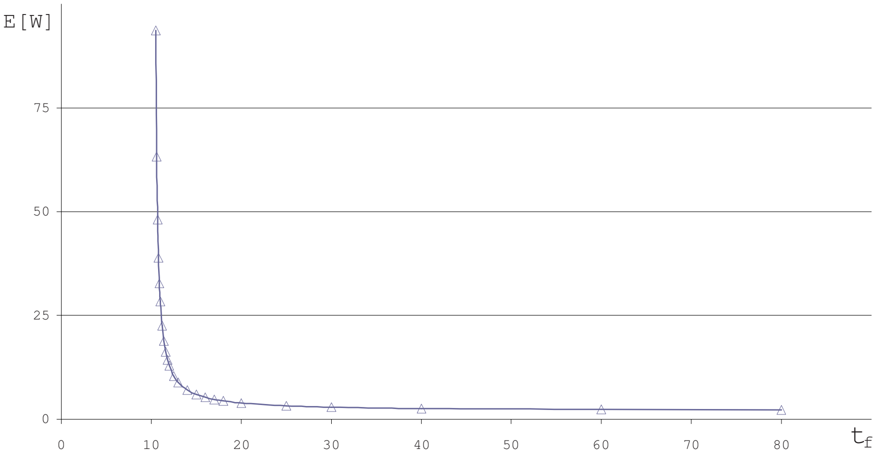


Fig. 18.5: Varying impact of interrupts (expressed in minutes) and the effect on patient waiting times (expressed in days)

among various sources of demand (e.g. patient classes). In a hospital setting this refers to the sharing of expensive diagnostic equipment, wards or labs. In a non-pooling environment, each resource fulfills its own demand, relying solely on its own capacity. In a pooled environment, demand is aggregated and fulfilled from a single shared facility. A rich literature on pooling in queueing systems exists. For an excellent overview, we refer to Benjaafar and Cooper (2005) and Yu and Benjaafar (2006).

It has long been known that pooling is beneficial to system performance. More specifically, pooling allows to maintain a specified level of service quality (e.g. patient flow times) with less capacity requirements. The beneficial effect of pooling stems from the increased ability of the system to cope with variability. For instance, in pooled systems, it is much less likely for the queue to be empty. As such, the impact of variability in the arrival pattern of patients (patients may arrive early, late or may even fail to show up at all) or in the service process of surgeons is minimized.

In this section, we demonstrate the impact of server pooling by means of a small experiment. We build on the setting of the hospital department discussed in the previous sections. In the experiment the servers at the consultation and surgery level are pooled. The following assumptions are imposed:

- Patients are treated by the first surgeon available for service, even if the patient does not belong to the patient population corresponding to that surgeon.
- Surgeon working schedules are identical and no structural constraints are imposed (i.e. it should be possible to service 6 patients simultaneously).

Returning to our example setting, the six consultation and the six surgery workstations are replaced by a single consultation and a single surgery workstation respectively. Each of these workstations has six parallel servers in operation. The resulting queueing network contains five workstations i, $i \in \{1,2,\ldots,5\}$. Let station 1 to 5 represent consultation, surgery, day hospital, internal ward and external ward respectively. When retaining all other characteristics of the setting discussed in the previous sections, one can use the multiserver Kingman equation to obtain patient flow times. The resulting performance measures are presented in Table 18.5 (the non-pooled flow times are the weighted average of the flow times observed at the consultation and surgery workstations presented in section 18.4.1).

The benefits of pooling are clear. Without increasing capacity or altering any of the other system characteristics (except of course the pooling of capacity) we are able to reduce patient flow times at the consultation and surgery level by a factor of 8.73 and 7.74 respectively.

Unfortunately, it is often impossible to achieve such a high degree of pooling in a real life hospital system. One quickly runs into a number of limitations:

- Unique relation between patient and surgeon; patients will often refuse to consult another surgeon.
- Limited flexibility of resources; each surgeon has his own specialization. It is often impossible, even for surgeons at the same department, to pass on jobs. In other words, the flexibility of surgeons is limited.

Table 18.5: Summary table of the model results after pooling (consultation and surgery workstations)

i	1	2
$\frac{1}{\mu_i}$	246.90	995.87
$C^2_{s_i}$	1.334	0.224
$\frac{1}{\lambda_i}$	43.56	173.2
$C^2_{a_i}$	0.996	1.075
A_i	0.101	0.079
ρ_i	0.944	0.958
$E[W_i]$ (pooled)	0.518	1.612
$E[W_i]$ (non-pooled)	4.523	12.47

- Resources often operate at different time instances; for pooling to take place surgeons need to operate at the same time instance. Due to busy schedules and other limitations, this is not always possible.
- Structural characteristics may further limit the practical applicability of pooling. For instance, if only two operating theatres are available, it is impossible to pool the capacity of the six surgeons at the surgery level. In other words, the bottleneck has shifted from the surgeons onto the number of available operating theatres.

Notwithstanding these constraints, it should be clear that even a small amount of pooling may yield significant reductions in patient flow time. Therefore the pooling of hospital resources is a worthwhile matter for further investigation.

18.4.4 Finding the optimal number of patients in a service session

The impact of absences at the start of a consultation or surgery session is discussed in Babes et al. (1991), Liu et al. (1998a), Liu and Liu (1998b) and Easton et al. (2005). There is a general agreement on the disruptive effect of absences on patient flow time. Easton et al. (2005) identify robust staffing, scheduling and recovery practices to minimize the effects of absences. Liu et al. (1998b) acknowledge the importance of consultation and surgery block size (i.e. the number of patients treated during a consultation session) and propose a what-if simulation approach in order to determine the best block size.

In fact, the relationship between block size and patient flow time is akin to the relationship between batch size and waiting time (in the presence of setups between batches in a manufacturing setting). As such the convex relationship first described by Karmarkar (1987) may also be observed here. In this view, Vandaele, Van Nieuwenhuyse and Cupers (2003b) determine the optimal size of patient groups queueing in front of a nuclear resonance scanner. We build on the model of Lambrecht and Vandaele (1996) in order to determine the optimal number of patients that receive treatment during a service session.

Two conflicting effects may be observed:

- The grouping effect; referring to the time required to assemble a batch of size n. The larger the batch size, the longer patients will have to wait before receiving service.
- The saturation effect; the smaller the batch size, the more service sessions are initiated, the larger the probability of having an absence of medical staff at the start of a service session.

We illustrate these effects in Figure 18.6. The combination of both effects results in a convex relationship, which implies that there is an optimal group size minimizing average patient lead time. In what follows, we develop the mathematical model to address the batch size decision problem. The objective is to determine the batch size that minimizes the average patient lead time.

In this section we build on the third workstation discussed in the base case (other workstations at the consultation and surgery level may also be analyzed in a similar fashion). To maintain the transparency of the model, we omit the index i referring to the original workstation used in this experiment. Other than the batching of patients, the dynamics of the workstation remain unchanged (as compared to previous sections).

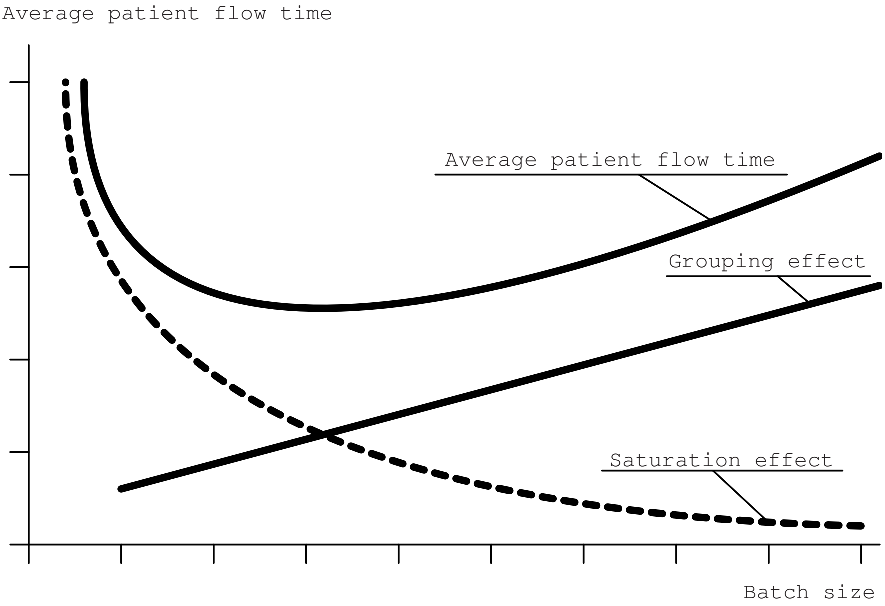


Fig. 18.6: Convex relationship between average patient flow time and batch size

Once sufficient patients are available, a batch (i.e. the equivalent of a service session workload) is created and is introduced into a queue (it is clear that this grouping does not imply that patients have to wait physically in the hospital). Whenever the server is idle, the batch as a whole receives service. After service, the batch is separated and patients resume their individual routings. A batch of patients is characterized by:

- a batch size n,
- a batch arrival rate λ_b,
- a coefficient of variation of the interarrival times of the batches $C^2_{a_b}$.
- a batch service rate μ_b,
- a coefficient of variation of the service times of the batches $C^2_{s_b}$,

where

$$\lambda_b = n\lambda, \tag{18.3}$$

$$C^2_{a_b} = \frac{C^2_a}{n}, \tag{18.4}$$

$$\mu_b = n\mu, \tag{18.5}$$

$$C^2_{s_b} = \frac{C^2_s}{n} \tag{18.6}$$

and λ, C^2_a, μ, C^2_s are the respective arrival rate, the squared coefficient of variation of the interarrival times, the service rate and the squared coefficient of variation of the service times of the individual patients visiting the third workstation.

The flow time of a patient in this system contains the following elements:

- The collection time; the time required until sufficient patients have arrived and a batch may be processed. The larger the batch size, the longer it takes to gather sufficient patients in order to perform a batch service.
- The waiting time of the batch itself; other batches (i.e. service sessions) may have to be serviced first.
- The absence time; prior to the service of a batch of patients, there exists a probability that the surgeon (or another crucial hospital resource) is absent. The batch of patients has to wait for the surgeon in order to receive service. This absence time can be considered as a setup time for the batch.
- The actual processing of individual patients in the batch.

We visualize the flow time of a patient in Figure 18.7. The expected flow time of a single patient in the system can be expressed as (Lambrecht and Vandaele, 1996):

$$E\left[W\right] = \frac{n-1}{2\lambda} + E\left[W_q\right] + \frac{1}{\mu_s} + \frac{n+1}{2\mu}. \tag{18.7}$$

This flow time clearly consists of four building blocks. The first term corresponds to the average time a patient will have to wait until a group of size n has been formed (i.e. the collection time). The term $E\left[W_q\right]$ stands for the average time that a batch of patients spends waiting in queue until the server becomes idle. We approximate $E[W_q]$ by means of the Kingman equation and obtain:

$$E[W_q] = \left(\frac{C_{a_b}^2 + C_{s_b}^2}{2}\right)\left(\frac{\rho}{1-\rho}\right)\frac{1}{\mu_b}, \tag{18.8}$$

where ρ is the effective utilization rate at the third workstation and is given by (Lambrecht et al., 1996):

$$\rho = \frac{n\lambda}{n\mu + \mu_s}. \tag{18.9}$$

The third term corresponds to the absence time that is incurred at the start of a service session in which a batch of patients receives treatment. Both the second and third term are the same for all patients in the batch. The last term indicates how much time a patient spends on processing itself. At this point the model is complete and we can formally state our optimization problem:

$$\begin{array}{ll} \text{Minimize}\, E[W], & E[W] = \frac{n-1}{2\lambda} + E[W_q] + \frac{1}{\mu_s} + \frac{n+1}{2\mu}, \\ s.t. & \rho < 1, \\ & n \geq 1. \end{array}$$

When using the setting of the hospital department outlined in the previous sections, we are able to provide a numerical example. To maintain transparency, we select a single consultation workstation and assess different values of n in order to obtain the optimal number of patients to be treated during a service session. A summary of the resulting figures is presented in Table 18.6.

An illustration is provided in Figure 18.8. One can deduce that, for this particular workstation, the optimum is reached when treating 8 patients during each service session. More precisely, given a set of input parameters (absence probability, service

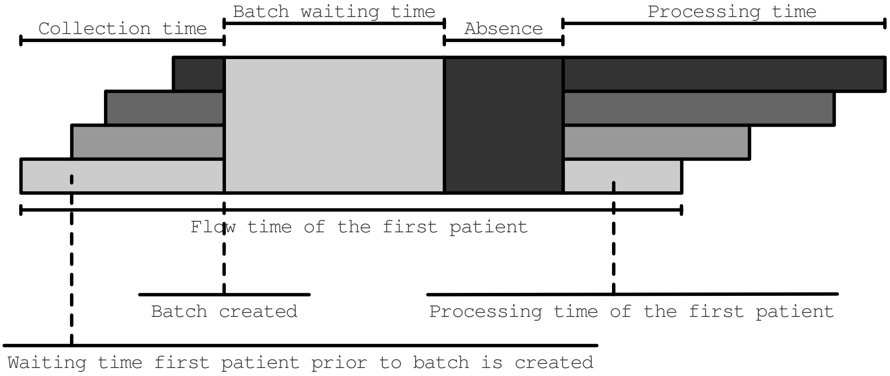

Collection time
Batch waiting time
Absence
Processing time
Flow time of the first patient
Batch created
Processing time of the first patient
Waiting time first patient prior to batch is created

Fig. 18.7: Visualization of the different phases of the batch flow time

Table 18.6: Summary table of the model results featuring different batch sizes

n	$\frac{1}{\mu_b}$	$C^2_{s_b}$	ρ	$E[W]$
3	82.063	0.2276	1.0787	NA
4	99.418	0.1707	0.9802	27.460
5	116.77	0.1365	0.9210	8.2226
6	134.13	0.1138	0.8815	6.3769
7	151.48	0.0975	0.8534	5.8782
8	168.84	0.0853	0.8322	5.7761
9	186.19	0.0758	0.8162	5.8441
10	203.54	0.0683	0.8027	6.0004
11	220.90	0.0621	0.7919	6.2086
12	238.25	0.0569	0.7830	6.4497
13	255.61	0.0525	0.7754	6.7132
14	272.96	0.0488	0.7689	6.9924
15	290.32	0.0455	0.7632	7.2831
16	307.67	0.0427	0.7583	7.5826
17	325.03	0.0402	0.7540	7.8888
18	342.38	0.0379	0.7501	8.2004
19	359.73	0.0359	0.7466	8.5162
20	377.09	0.0341	0.7435	8.8355

and interarrival times, . . .) we are able to determine the optimal number of patients to be treated during a service session.

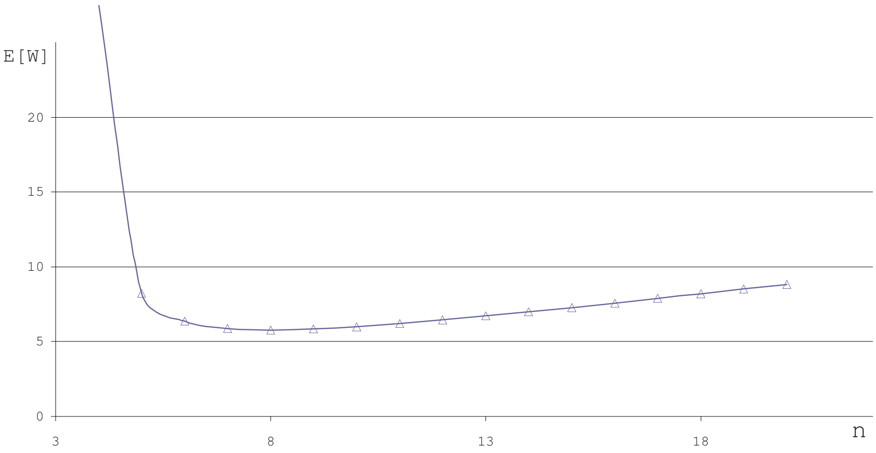


Fig. 18.8: Finding the optimal number of patients

18.5 Conclusion

In this article we discuss some of the features that differ when modeling healthcare queueing models on the one hand and traditional manufacturing models on the other hand. We show how to implement these features in a hospital queueing network. We used the parametric decomposition approach to assess performance measures at the hospital queueing network. In addition, we develop new expressions to model service outages that are typical in services in general and in healthcare in particular. The resulting queueing network is used to construct a numerical example and to illustrate a number of practical applications. First we demonstrate the detrimental effect of service interrupts on patient flow times. Next, the beneficial effect of pooling hospital resources is illustrated. Finally, we develop an optimization model that is able to determine the optimal number of patients treated during a single service session.

Notwithstanding these accomplishments, there is still room for improvement. More specifically, improvements may be made with respect to the modeling of time in queueing systems. Open problems include the modeling of time-dependent demand rates, increasing workload as waiting times increase (patients need to be monitored, receive care, . . .), ... Moreover, given the inherent high degree of variability in service times, hospitals often use flexible working schedules that allow for overtime, variable server capacity and other deviations from the standard queueing model topology. Such deviations add to the complexity of the problem, making “time” a major modeling issue.

References

1. Askin RG (1993) Modeling and analysis of manufacturing systems. Wiley, New York
2. Albin SL (1984) Approximating a point process by a renewal process, II: superposition arrival processes to queues. Operations Research 32:1133–1162
3. Babes M, Sarma GV (1991) Out-patient queues at the Ibn-Rochd health center. Journal of the Operational Research Society 42:845–855
4. Benjaafar S, Cooper WL (2005) On the benefits of pooling in production-inventory systems. Management Science 51:548–565
5. Bitran GR, Tirupati D (1988) Multiproduct queueing networks with deterministic routing: decomposition approach and the notion of interference. Management Science 34:75–100
6. Buzacott JA, Shanthikumar JG (1985) Queueing Models of Dynamic Job Shops. Management Science 31:870–887
7. Chisholm CD, Collison EK, Nelson DR, Cordell WH (2000) Emergency department workplace interruptions: are emergency physicians “interrupt-driven” and “multitasking”. Academic Emergency Medicine 7:1239–1243
8. Chisholm CD, Dornfeld AM, Nelson DR, Cordell WH (2001) Work interrupted: a comparison of workplace interruptions in emergency departments and primary care offices. Annals of Emergency Medicine 38:146–151
9. Creemers S, Lambrecht MR (2007) Modeling a healthcare system as a queueing network: the case of a Belgian hospital. In: FBE publications: Research Reports and Discussion papers. Department of Decision Sciences & Information Management, Research Center for Opera-

tions Management, Catholic University Leuven. Available via KULeuven. http://bib.kuleuven.be/ebib/wp.htm.Cited1Aug2008

10. Doshi BT (1986) Queueing systems with vacations - a survey. Queueing Systems 1:29–66
11. Dudewicz EJ, Mishra SN (1988) Modern mathematical statistics. John Wiley Sons, New York
12. Easton FF, Goodale JC (2005) Schedule recovery: unplanned absences in service operations. Decision Sciences 36:459–488
13. France DK, Levin S, Hemphill R, Chen K, Rickard D, Makowski R, Jones I, Aronsky D (2005) Emergency physicians' behaviors and workload in the precense of an electronic whiteboard. International Journal of Medical Informatics 74:827–837
14. Gabow PA, Karkhanis A, Knight A, Dixon P, Eiser S, Albert RK (2006) Observations of residents' work activities for 24 consecutive hours: implications for workflow redesign. Academic Medecine 81:766–775
15. Green LV, Soares J (2007) Computing Time-Dependent Waiting Time Probabilities in M(t)/M/s(t) Queueing Systems. M&SOM Manufacturing & Service Operations Management 9:54–61
16. Hall RW (2006) Patient Flow: The New Queueing Theory for healthcare. OR/MS Today 23:36–40
17. Hall RW, Belson D, Muralli P, Dessouky M (2006) Modeling patient flows through the healthcare system. In: Hall RW (ed) Patient flow: reducing delay in healthcare delivery, Springer Science, New York
18. Haskose A, Kingsman BG, Worthington D (2002) Modelling flow and jobbing shops as a queueing network for workload control. International Journal of Production Economics 78:271–285
19. Haque L, Armstrong MJ (2007) A survey of the machine interference problem. European Journal of Operational Research 179:469–482
20. Harvey R, Jarrett PG, Peltekian KM (1994) Patterns of paging medical interns during night calls at two teaching hospitals. Canadian Medical Association Journal 151:307–311
21. Hopp WJ, Spearman L (2000) Factory Physics. McGraw-Hill Higher Education, New York
22. Jackson JR (1957) Network of waiting lines. Operations Research 5:518–521
23. Jackson JR (1963) Jobshop-like queueing systems. Management Science 10:131–142
24. Karmarkar US (1987) Lot sizes, lead times and in-process inventories. Management Science 33:409–418
25. Khinchin AJ (1960) Mathematical Methods in the Theory of Queueing. Hafner, New York
26. Kingman JFC (1962) On queues in heavy traffic. Journal of the Royal Statistical Society. Series B (Methodological) 24:383–392
27. Lambrecht MR, Vandaele NJ (1996) A general approximation for the single product lot sizing model with queuing delays. European Journal of Operational Research 95:73–88
28. Lambrecht MR, Ivens PL, Vandaele NJ (1998) ACLIPS: a capacity and lead time integrated procedure for scheduling. Management Science 44:1548–1561
29. Lariviere MA, Van Mieghem JA (2004) Strategically seeking service: how competititon can generate Poisson arrivals. Manufacturing & Service Operations Management 6:23–40
30. Lehaney B, Clarke SA, Paul RJ (1999) A case of intervention in an outpatient department. Journal of the Operational Research Society 50:877–891
31. Liu L, Liu X (1998a) Block appointment systems for outpatient clinics with multiple doctors. The Journal of the Operational Research Society 49:1254–1259
32. Liu L, Liu X (1998b) Dynamic and static job allocation for multi-server systems. IIE Transactions 30:845–854
33. Marshall KT (1968) Some inequalities in queuing. Operations Research 16:651–668
34. Palm C (1943) Intensittsschwankungen im Fernsprechverkehr. Ericsson Technics 44:1–89
35. Roth A, Van Dierdonck R (1995) Hospital resource planning: concepts, feasibility and framework. Production and Operations Meanagement 4:2–29
36. Shanthikumar JG, Buzacott JA (1981) Open queueing network models of dynamic job shops. International Journal of Production Research 19:255–266

37. Sethuraman K, Tirupati D (2005) Evidence of bullwhip effect in healthcare sector: causes, consequences and cures. International Journal of Services and Operations Management 1:372–394
38. Stecke KE, Aronson JE (1985) Review of operator/machine interference models. Journal of Production Research 23:129–151
39. Suri R, Sanders JL, Kamath M (1993) Performance evaluation of production networks. In: Graves SC et al. (ed) Handbooks in Operations Research and Management Science, Vol. 4: Logistics of Production and Inventory, Elsevier Science Publishers, New York
40. Takagi H (1988) Queueing analysis of polling models. ACM Computing Surveys 20:5–28
41. Tian N, Zhang ZG (2006) Vacation queueing models. Springer Science, New York
42. Tucker AL, Spear SJ (2006) Operational failures and interruptions in hospital nursing. Health Services Research 41:643–662
43. van Merode GG, Groothuis S, Hasman A (2004) Enterprise resource planning for hospitals. International Journal of Medical Informatics 73:493–501
44. Vandaele N, De Boeck L (2003a) Advanced resource planning. Robotics and Computer Integrated Manufacturing 19:211–218
45. Vandaele N, Van Nieuwenhuyse I, Cupers S (2003b) Optimal grouping for a nuclear magnetic resonance scanner by means of an open queueing model. European Journal of Operational Research 151:181–192
46. Vissers JMH, Bertrand JWM, De Vries G (2001) A framework for production control in health care organizations. Production Planning & Control 12:591–604
47. Volpp KGM, Grande D (2006) Residents' suggestions for reducing errors in teaching hospitals. The New England Journal of Medicine 348:851–855
48. Whitt W (1983) The queueing network analyzer. The Bell System Technical Journal 62:2779–2815
49. Whitt W (1994) Towards better multi-class parametric-decomposition approximations for open queueing networks. Annals of Operations Research 48:221–248
50. Whitt W (1995) Variability functions for parametric-decomposition approximations of queueing networks. Management science 41:1704–1715
51. Whitt W (1999a) Decomposition approximations for time-dependent Markovian queueing networks. Operations Research Letters 24:97–103
52. Whitt W (1999b) Partitioning Customers into Service Groups. Management Science 45:1579–1592
53. Yu Y, Benjaafar S, Gerchak Y (2006) On service capacity pooling and cost sharing among independent firms. Department of Mechanical Engineering, University of Minnesota. Available via University of Minnesota.
http://www.ie.umn.edu/faculty/faculty/pdf/ybg06.pdf.
Cited1Aug2008