GENERAL QUEUEING NETWORK MODELS
FOR COMPUTER SYSTEM PERFORMANCE ANALYSIS

A maximum entropy method of analysis and aggregation of general queueing network models with application to computer systems

Mohamed Ahmed El-Affendi
PhD
1983

Postgraduate School of Studies in Computing
University of Bradford
ABSTRACT

In this study the maximum entropy formalism [JAYN 57] is suggested as an alternative theory for general queueing systems of computer performance analysis. The motivation is to overcome some of the problems arising in this field and to extend the scope of the results derived in the context of Markovian queueing theory.

For the M/G/1 model a unique maximum entropy solution, satisfying local balance is derived independent of any assumptions about the service time distribution. However, it is shown that this solution is identical to the steady state solution of the underlying Markov process when the service time distribution is of the generalised exponential (GE) type. (The GE-type distribution is a mixture of an exponential term and a unit impulse function at the origin). For the G/M/1 the maximum entropy solution is identical in form to that of the underlying Markov process, but a GE-type distribution still produces the maximum overall similar distributions.

For the G/G/1 model there are three main achievements:
(i) first, the spectral methods are extended to give exact formulae for the average number of customers in the system for any G/G/1 with rational Laplace transform. Previously, these results are obtainable only through simulation and approximation methods.
(ii) secondly, a maximum entropy model is developed and used to obtain unique solutions for some types of the G/G/1. It is also discussed how these solutions can be related to the corresponding stochastic processes.
(iii) the importance of the G/GE/1 and the GE/GE/1 for the analysis of general networks is discussed and some flow processes for these systems are characterised.
For general queueing networks it is shown that the maximum entropy solution is a product of the maximum entropy solutions of the individual nodes. Accordingly, existing computational algorithms are extended to cover general networks with FCFS disciplines. Some implementations are suggested and a flow algorithm is derived. Finally, these results are used to improve existing aggregation methods.

In addition, the study includes a number of examples, comparisons, surveys, useful comments and conclusions.

DECLARATION

Chapters III, IV, V and VI describe work done in conjunction and close collaboration with my supervisor, Dr D D Kouvatsos. These chapters are in essence transcripts of joint research papers which have been submitted or accepted for publication (see References ELAF, 82, 83a, 83b).
ACKNOWLEDGEMENTS

I am very grateful to Dr D D Kouvatsos for his close supervision, continued support and exceptional enthusiasm. He has been very helpful, encouraging and greatly understanding throughout the development of this study.

I would also like to thank my parents for their care, support and encouragement. I am also deeply grateful to my wife for her patience, support and care.

I would also like to thank Mrs D F Connor for her professional contribution in typing this thesis.

I am also grateful to the University of Khartoum for sponsoring this research.
# CONTENTS

<table>
<thead>
<tr>
<th>Page No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
</tr>
<tr>
<td>Acknowledgements</td>
</tr>
<tr>
<td>Contents</td>
</tr>
<tr>
<td>List of Tables</td>
</tr>
<tr>
<td>List of Figures</td>
</tr>
<tr>
<td>Basic Notation</td>
</tr>
</tbody>
</table>

## CHAPTER I  GENERAL INTRODUCTION

8

## CHAPTER II  THE MAXIMUM ENTROPY FORMALISM

2.1 An Overview 8

2.2 Basic Properties of the Entropy Functionals 12

2.3 The Maximum Entropy Formalism 15

## CHAPTER III  MAXIMUM ENTROPY ANALYSIS OF THE M/G/1 AND THE G/M/1 QUEUEING SYSTEMS IN EQUILIBRIUM

3.1 Introduction 19

3.2 A Maximum Entropy Solution for the M/G/1

3.2.1 The constraints of the M/G/1 20

3.2.2 The maximum entropy model 21

3.2.3 The underlying service time distribution 23

3.3 Maximum Entropy Formalism and the G/M/1

3.3.1 Statement of the problem 27

3.3.2 A maximum entropy solution 28

3.4 Balance Equations for the Maximum Entropy M/G/1-G/M/1 Systems 32
### CHAPTER IV SPECTRAL METHODS AND MAXIMUM ENTROPY IN THE ANALYSIS OF THE G/G/1 QUEUEING SYSTEM

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Introduction</td>
<td>51</td>
</tr>
<tr>
<td>4.2</td>
<td>Exact Results for Some Variations of the G/G/1</td>
<td>52</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Fundamental queueing theory results</td>
<td>53</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Analysis of the G/GE/1 queueing system</td>
<td>55</td>
</tr>
<tr>
<td>4.2.3</td>
<td>The average number of customers in the GE/G/1</td>
<td>66</td>
</tr>
<tr>
<td>4.2.4</td>
<td>Other variations of the G/G/1 system</td>
<td>71</td>
</tr>
<tr>
<td>4.2.5</td>
<td>Numerical comparisons</td>
<td>75</td>
</tr>
<tr>
<td>4.3</td>
<td>Maximum Entropy and the G/G/1</td>
<td>83</td>
</tr>
<tr>
<td>4.3.1</td>
<td>A maximum entropy model for general single queues</td>
<td>84</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Maximum entropy and the G/GE/1</td>
<td>86</td>
</tr>
<tr>
<td>4.4</td>
<td>Maximum Entropy and Operational Analysis</td>
<td>91</td>
</tr>
<tr>
<td>4.4.1</td>
<td>Operational assumptions, laws and theorems</td>
<td>91</td>
</tr>
<tr>
<td>4.4.2</td>
<td>Maximum entropy queue recursions</td>
<td>94</td>
</tr>
<tr>
<td>4.5</td>
<td>Flow in the G/GE/1 and the GE/GE/1</td>
<td>97</td>
</tr>
<tr>
<td>4.5.1</td>
<td>The idle time distribution and the departure process</td>
<td>97</td>
</tr>
<tr>
<td>4.5.2</td>
<td>The GE/GE/1 with Bernoulli feedback</td>
<td>104</td>
</tr>
<tr>
<td>4.6</td>
<td>Summary</td>
<td>107</td>
</tr>
</tbody>
</table>
CHAPTER V TWO-SERVER TANDEM AND CYCLIC QUEUES WITH NON-EXPONENTIAL SERVICE TIMES

5.1 Introduction

5.2 The Maximum Entropy G/G/1 with Finite Waiting Room

5.3 Analysis of Tandem and Cyclic Queues
   5.3.1 A general two-station tandem queueing system
   5.3.2 A general two-station cyclic queueing model

5.4 Correlation in the Departure Process

5.5 Summary

CHAPTER VI MAXIMUM ENTROPY ANALYSIS OF GENERAL QUEUEING NETWORKS

6.1 Introduction

6.2 Maximum Entropy and General Queueing Networks
   6.2.1 Description of the model
   6.2.2 Maximum entropy model

6.3 Implementation of the Maximum Entropy Model
   6.3.1 Maximum entropy and Markovian queueing analysis
   6.3.2 Flow in the network
   6.3.3 Maximum entropy and operational analysis

6.4 Examples and Applications

6.5 Summary

CHAPTER VII AGGREGATION OF GENERAL QUEUEING NETWORKS

7.1 Introduction

7.2 An Overview
   7.2.1 The near-complete-decomposability approach of Courtois [COUR 75,77]
   7.2.2 The Norton's theorem approach
7.3 Problems of Aggregation

7.4 Contribution of the Maximum Entropy Analysis
   7.4.1 Modification of the Courtois algorithm
   7.4.2 Modification of the flow equivalent algorithm

7.5 Comparisons and Applications

7.6 Summary

CHAPTER VIII CONCLUSIONS AND FUTURE PROSPECTS

8.1 General Summary

8.2 Advantages of the Analysis

8.3 Future Prospects

REFERENCES
LIST OF TABLES

3.1 The average number of customers in a GE/M/1 compared to that of an equivalent E\textsubscript{2}/M/1 and the diffusion approximation

3.2 Parameters of the maximum entropy M/G/1 and G/M/1 systems

4.1 The average number of customers in the system compared for some variations of the G/G/1: the E\textsubscript{2}/H\textsubscript{2}/1 versus simulation and other systems

4.2 The average number of customers compared for some types of the G/G/1: the E\textsubscript{2}/H\textsubscript{2}/1 versus other systems

4.3 The average number of customers compared for some types of the G/G/1: the E\textsubscript{2}/H\textsubscript{2}/1 versus other systems

4.4 The average number of customers compared for some types of the G/G/1: the H\textsubscript{2}/H\textsubscript{2}/1 versus other systems

4.5 The average number of customers compared for some types of the G/G/1: the E\textsubscript{2}/Hypo\textsubscript{2}/1 versus other systems

4.6 Summary of the results obtained using spectral methods (section 4.3)

5.1 The average number of customers and the mean waiting time for station 2 in the tandem queue of example 5.3.1(i)

5.2 The average number of customers and the mean waiting time for station 2 in the tandem queue of example 5.3.1(ii)

5.3 The average number of customers and the mean waiting time in the tandem queue of example 5.3.1(iii)

5.4a The utilisation of station 2 in the cyclic model of example 5.3.1(i)

5.4b The average number of customers for station 2 in the cyclic model of example 5.3.1(i)

5.5a The utilisation of station 2 in the cyclic model of example 5.3.2(ii)

5.5b The average number of customers for station 2 in the cyclic model of example 5.3.2(ii)

5.6 The utilisation and the average number of customers in the M/H\textsubscript{2}/1/N for different values of k in the expression for \(a_1\) (3.41b).
5.7 Correlation in the departure process for an M/GE/1

6.1 The average number of customers at each node i of the open network compared with simulation, example 6.4.1(i).

6.2 The average number of customers at each node of the open network of example 6.4.1(ii), compared with simulation and the diffusion approximation.

6.3 The average number of customers at each node i of the open network compared with simulation, example 6.4.1(iii).

6.4 Some performance metrics for the central server model of Fig. 6.4.

6.5 Some performance metrics for the central server model of Fig. 6.5.

6.6 Flow parameters for the models of Fig. 6.4 and Fig. 6.5.

7.1 Performance metrics for the model in Figs. 7.8-7.10.
<table>
<thead>
<tr>
<th>Fig.</th>
<th>Description</th>
<th>Page No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1a</td>
<td>Balance diagram for the M/G/1</td>
<td>34</td>
</tr>
<tr>
<td>3.1b</td>
<td>Balance diagram for the G/M/1</td>
<td>34</td>
</tr>
<tr>
<td>4.1</td>
<td>The contour C for the G/GE/1</td>
<td>58</td>
</tr>
<tr>
<td>4.2</td>
<td>The excursion around the origin</td>
<td>58</td>
</tr>
<tr>
<td>4.3</td>
<td>Behaviour sequence for a queue for a 10-second observation period</td>
<td>92</td>
</tr>
<tr>
<td>4.4</td>
<td>A queue with a feedback</td>
<td>105</td>
</tr>
<tr>
<td>5.1</td>
<td>A tandem queueing system</td>
<td>110</td>
</tr>
<tr>
<td>5.2a</td>
<td>An open cyclic queueing model</td>
<td>110</td>
</tr>
<tr>
<td>5.2b</td>
<td>A closed cyclic queueing model</td>
<td>110</td>
</tr>
<tr>
<td>5.3</td>
<td>A two-station tandem queue, example 5.3.1(i-iii)</td>
<td>116</td>
</tr>
<tr>
<td>5.4</td>
<td>A two-station cyclic queueing model, example 5.3.2(i,ii)</td>
<td>116</td>
</tr>
<tr>
<td>6.1</td>
<td>An example of a general queueing network</td>
<td>139</td>
</tr>
<tr>
<td>6.2</td>
<td>Merging two streams</td>
<td>151</td>
</tr>
<tr>
<td>6.3</td>
<td>The open network of example 6.4.1</td>
<td>157</td>
</tr>
<tr>
<td>6.4</td>
<td>Example 6.4.2, model 1</td>
<td>162</td>
</tr>
<tr>
<td>6.5</td>
<td>Example 6.4.2, model 2</td>
<td>162</td>
</tr>
<tr>
<td>7.1</td>
<td>A central server model</td>
<td>171</td>
</tr>
<tr>
<td>7.2</td>
<td>The two-server model equivalent to the central server network of Fig. 7.1</td>
<td>171</td>
</tr>
<tr>
<td>7.3</td>
<td>The procedure of multilevel aggregation</td>
<td>173</td>
</tr>
<tr>
<td>7.4</td>
<td>The transition matrix Q for the system in Fig. 7.1</td>
<td>174</td>
</tr>
<tr>
<td>7.5</td>
<td>The original network (flow-equivalent method)</td>
<td>179</td>
</tr>
<tr>
<td>7.6</td>
<td>The equivalent subnetwork SUBQi with node i replaced by a short</td>
<td>179</td>
</tr>
</tbody>
</table>
Fig. 7.7 The equivalent two-server model 180
Fig. 7.8 Model 1 of the example in section 7.5.2 187
Fig. 7.9 Model 2 of the example in section 7.5.2 187
Fig. 7.10 Model 3 of the example in section 7.5.2 188
### BASIC NOTATION

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A(t)$</td>
<td>the PDF of interarrival times</td>
</tr>
<tr>
<td>$a(t)$</td>
<td>the pdf of interarrival times</td>
</tr>
<tr>
<td>$A^*(\theta)$</td>
<td>the Laplace transform of $a(t)$</td>
</tr>
<tr>
<td>$\mu_a$</td>
<td>the first moment of $a(t)$</td>
</tr>
<tr>
<td>$\sigma^2_a$</td>
<td>the second moment of $a(t)$</td>
</tr>
<tr>
<td>$C_a^2$</td>
<td>the coefficient of variation of interarrival times</td>
</tr>
<tr>
<td>$C_d^2$</td>
<td>the coefficient of variation of interdeparture times</td>
</tr>
<tr>
<td>$C_s^2$</td>
<td>the coefficient of variation of service times</td>
</tr>
<tr>
<td>$D(t)$</td>
<td>the PDF of the interdeparture times</td>
</tr>
<tr>
<td>$d(t)$</td>
<td>the pdf of the interdeparture times</td>
</tr>
<tr>
<td>$D^*(\theta)$</td>
<td>the Laplace transform of $d(t)$</td>
</tr>
<tr>
<td>$\mu_D$</td>
<td>the first moment of $d(t)$</td>
</tr>
<tr>
<td>$\sigma^2_D$</td>
<td>the second moment of $d(t)$</td>
</tr>
<tr>
<td>$E_k$</td>
<td>Erlang$_k$ distribution</td>
</tr>
<tr>
<td>$E(x)$</td>
<td>Expected value of the random variable $x$</td>
</tr>
<tr>
<td>$F(t)$</td>
<td>the PDF of service times</td>
</tr>
<tr>
<td>$f(t)$</td>
<td>the pdf of service times</td>
</tr>
<tr>
<td>$F^*(\theta)$</td>
<td>the Laplace transform of $f(t)$</td>
</tr>
<tr>
<td>$\mu_f$</td>
<td>the first moment of $f(t)$</td>
</tr>
<tr>
<td>$\sigma^2_f$</td>
<td>the second moment of $f(t)$</td>
</tr>
<tr>
<td>FCFS</td>
<td>first-come, first-served</td>
</tr>
<tr>
<td>$\text{GE}$</td>
<td>the generalised exponential distribution</td>
</tr>
<tr>
<td>$H_2$</td>
<td>Hyperexponential$_2$ distribution</td>
</tr>
<tr>
<td>$H(p)$</td>
<td>the system entropy</td>
</tr>
<tr>
<td>$H(p,g)$</td>
<td>the system relative entropy ($g = $ prior distribution)</td>
</tr>
<tr>
<td>Symbol</td>
<td>Definition</td>
</tr>
<tr>
<td>------------</td>
<td>---------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Hypo\textsubscript{2}</td>
<td>hypoexponential\textsubscript{2} distribution</td>
</tr>
<tr>
<td>I(t)</td>
<td>the PDF of the idle time distribution</td>
</tr>
<tr>
<td>i(t)</td>
<td>the pdf of the idle time distribution</td>
</tr>
<tr>
<td>I*(\theta)</td>
<td>the Laplace transform of the idle time distribution</td>
</tr>
<tr>
<td>\bar{I}</td>
<td>the first moment of i(t)</td>
</tr>
<tr>
<td>\bar{I}^2</td>
<td>the second moment of i(t)</td>
</tr>
<tr>
<td>&lt;n&gt;</td>
<td>the average number of customers in the system</td>
</tr>
<tr>
<td>PDF</td>
<td>probability distribution function</td>
</tr>
<tr>
<td>pdf</td>
<td>probability density function</td>
</tr>
<tr>
<td>P\textsubscript{n}</td>
<td>the outside observer's probability distribution that</td>
</tr>
<tr>
<td></td>
<td>there are n customers in the system</td>
</tr>
<tr>
<td>T</td>
<td>the average response time</td>
</tr>
<tr>
<td>W(t)</td>
<td>the PDF of the waiting time distribution</td>
</tr>
<tr>
<td>w(t)</td>
<td>the pdf of the waiting time distribution</td>
</tr>
<tr>
<td>W*(\theta)</td>
<td>the Laplace transform of w(t)</td>
</tr>
<tr>
<td>W</td>
<td>the mean waiting time</td>
</tr>
<tr>
<td>Z\textsubscript{p}</td>
<td>the partition function</td>
</tr>
<tr>
<td>\Pi</td>
<td>product sign</td>
</tr>
<tr>
<td>\lambda</td>
<td>mean arrival rate</td>
</tr>
<tr>
<td>\mu</td>
<td>mean service rate</td>
</tr>
<tr>
<td>P\textsubscript{n}</td>
<td>the arriver's probability distribution that</td>
</tr>
<tr>
<td></td>
<td>there are n customers in the system</td>
</tr>
<tr>
<td>P\textsubscript{n}'</td>
<td>the completer's probability distribution that</td>
</tr>
<tr>
<td></td>
<td>there are n customers in the system</td>
</tr>
<tr>
<td>\rho</td>
<td>\lambda/\mu the utilisation factor</td>
</tr>
<tr>
<td>\Sigma</td>
<td>summation sign</td>
</tr>
<tr>
<td>\sigma\textsuperscript{2} \textsubscript{a}</td>
<td>variance of interarrival times</td>
</tr>
<tr>
<td>\sigma\textsuperscript{2} \textsubscript{f}</td>
<td>variance of service times</td>
</tr>
</tbody>
</table>
CHAPTER I
GENERAL INTRODUCTION

Queueing network models are now widely applied in the performance analysis of computer systems. These models were first introduced by Jackson [JACK 57], where he considered open queueing networks with exponential service times. In a later publication, Jackson [JACK 63] generalised his results to queueing networks with feedback, closed chains, open chains, and load-dependent service rates. The special case of closed queueing networks was separately considered by Gordon and Newell [GORD 67].

Although Jackson's theory was introduced in 1957, application to computer systems did not effectively start until 1971, when Buzen [BUZE 71] and Moore [MOOR 72] produced the first applications of Jacksonian networks to computer systems. The only related successful application that preceded the work of Moore and Buzen was due to Scherr [SCHE 67] where he applied a machine repairman model to analyse a computer system. These early applications revealed how successful queueing networks can be in predicting the performance metrics of computer systems. However, it was realised at the same time that there is some computational complexity associated with closed queueing networks. This problem was greatly overcome when Buzen [BUZE 73] introduced an efficient computational algorithm for closed queueing networks.

The appearance of the Buzen algorithm aroused interest in queueing networks and led to a new era in the history of computer systems performance evaluation. During the period 1973-1976 a great deal of effort
was exerted to extend and generalise queueing network models. The culmination of this effort was the appearance of the BCMP* theorem by Baskett et al [BASK 75] which extends the results of Jackson to include different queueing disciplines, multiple classes of jobs, and non-exponential service distributions. Queueing networks that satisfy the BCMP theorem are generally known in the literature as Markovian queueing networks or locally balanced networks.

Despite persistent attempts for generalisation, three major problems remained unresolved:

(i) Markovian queueing theory could not provide a solution for queueing networks with non-exponential service times and FCFS (first come, first served) queueing disciplines. Even for the single resource model with arbitrary arrivals and service times, no exact solution can be obtained if the service discipline is FCFS [KLEI 75].

(ii) Even for the simplest Jacksonian network, it has been shown by Burke [BURK 76] that if feedback is allowed then the flow in the network is no longer Poisson. Disney and McNickle [DISN 76] further showed that in such a case the flow is not even a renewal process. This implies that in the case of feedback interinput and interdeparture times are correlated and Markovian queueing theory can no longer apply.

(iii) Markovian queueing theory could not give an explanation as to why queueing network models are so robust. Despite serious violation of the underlying assumptions, extensive experiments (see the references in [DENN 78]) revealed that the obtained performance metrics are remarkably accurate.

* BCMP stands for Baskett-Chandy-Muntz-Palacios
In response to these problems, two schools of thought emerged:
The first school represents analysts who are content with Markovian
queueing analysis and believe that stochastic theory represents the
best framework for performance analysis. The only course open for
this camp was to identify the best approximations for the actual
solutions of the underlying stochastic process. Most of the work
done in this respect was directed towards problem (i) above and
comprises two approximation techniques: decomposition and the
diffusion approximation. The pioneers of decomposition are Chandy et al
\cite{CHAN75} who introduced the Norton theorem approach for the decom-
position of queueing networks and Courtois \cite{COUR75} who developed the near
complete decomposability approach. Diffusion approximation methods
for queueing networks were developed by Gelenbe \cite{GELE74}, Reiser and
Kobayashi \cite{REIS74}, Gelenbe and Pujolle \cite{GELE75}.

To validate these approximations, simulation methods are often
used. In this respect the regenerative method for simulation,
developed by Iglehart and Schedler \cite{ICLE78} has been very useful.

The other school represents analysts who are less satisfied with
Markovian queueing analysis and believe that an alternative theory
should be sought. For them, stochastic analysis involves unverifiable
assumptions and incurs an unnecessary complexity. These views were
expressed by Denning and Buzen \cite{DENN78} and more strongly by Spragins
\cite{SPRA80}, who stated "An unfortunate percentage of the literature on
performance modelling of complex systems satisfies Rosonoff's definition
of durable nonsense: an alloy of sense and irrelevancy, protected with
a thick coating of rigour and abstraction, prepackaged in a convenient
black box, produce the most durable nonsense known to man." This trend
of thinking led to the appearance of operational analysis as an alternative to Markovian queueing analysis [DENN 78|BUZE 80].

In operational analysis, stochastic assumptions are replaced by testable operational assumptions on the basis of which operational variables and laws are defined. Operational laws lead to the same solutions obtained using exponential Markovian queueing networks, but without the need to assume any model for the service time distribution. In this way operational analysis avoids problems (i) and (ii) above and gives a possible explanation for (iii). Moreover, it greatly simplifies the analysis. For these reasons operational analysis has become popular among practitioners in the field.

Although operational analysis is an important step forward in reformulating the queueing problem, the underlying theory still suffers from a number of restrictions. Firstly, operational assumptions and laws hold only for a limited observation period. For this reason operational analysis relies heavily on validation and verification experiments. This dependence on empirical verification implicitly renounces the possibility for prediction. In the words of Jaynes [JAYN 57b], "Whenever we use a density matrix whose probabilities are verifiable by certain measurements, we necessarily renounce the possibility of predicting the results of other measurements which can be made on the same apparatus. This principle of statistical complementarity is not restricted to quantum mechanics but holds in any application of probability theory."

Secondly, operational analysis results do not take into account the service and interarrival time coefficients of variation. The effect of excluding these parameters is not yet known.
Thirdly, some of the operational assumptions may be necessary, but not sufficient for the operational theorems e.g. the queue recursions discussed by Buzen and Denning [BUZE 80] may hold in a wider sense without the need for the one-step transition assumption.

Therefore, while stochastic theory is highly abstract and complex, operational analysis goes for the other extreme and makes the analysis highly dependable on measurements and empirical verification. This, and the above problems of Markovian analysis stresses the need for a new universal theory that serves as a suitable framework for queueing network models of computer system performance.

To be successful, this theory should satisfy the following requirements:

a) give a convincing interpretation for a probability assignment.

b) eliminate the need for arbitrary assumptions.

c) simplify the analysis.

d) solve or eliminate the existing problems.

e) include the existing solutions as a subset.

f) explain the robustness of queueing networks.

The interpretation of a probability assignment has been the subject of a prolonged discussion between two schools of thought in probability theory: the objective school who advocate a frequency interpretation of probability and the subjective school for whom a probability assignment is not a property of any system but only a means of describing our information about the system. Jaynes [JAYN 57b] discussed this problem and concluded that if our models are to be used for prediction, then the subjective interpretation is favoured.
Elimination of arbitrary assumptions has been the motivation behind the principle of insufficient reason given long ago by Bernoulli which implies that:

"(1) a probability assignment is a state of knowledge,

(2) in determining a probability distribution the outcomes of an event should be considered initially equally probable unless there is evidence to make us think otherwise."

Although the principle of insufficient reason existed for more than a century, its significance has largely been ignored until the recent work of E T Jaynes [JAYN 57a, 57b], who extended Shannon's theory [SHAN 48] to give an objective mathematical meaning for the principle of insufficient reason. Jaynes named his theory "The Maximum Entropy Formalism". This formalism is an important step towards a unified probability theory since it "objectivises" most of the rules in subjective probability theory.

Moreover, applications of the maximum entropy formalism in statistical mechanics [SHOR 81] and spectral analysis revealed that it greatly simplifies the analysis and all the time the analyst seems to be getting something for nothing.

Based on these points it appears that the maximum entropy formalism can be a reasonable theoretical framework for the performance analysis of computer and information systems in general.

Maximum entropy can be applied in two ways:

1. Directly in conjunction with the available empirical data. The application in this case will be similar to operational analysis.

2. As a complementary theory to stochastic analysis in which case it will refine the results and reduce the number of arbitrary assumptions.
If the model is to be used for consistency checking or validation, then the first approach is recommended. However, if the model is to be used for prediction, then the second approach may be favourable.

This study is an attempt to introduce maximum entropy as a theoretical framework for queueing networks of computer system performance. Maximum entropy is predominantly applied to extend and rectify the existing Markovian queueing theory results.

In Chapter 2 the maximum entropy formalism is introduced. The M/G/1 and the G/M/1 systems are considered in Chapter 3 in an attempt to illustrate the advantages of the maximum entropy formalism. Chapter 4 represents the core of the study where spectral methods and maximum entropy play a joint role in analysing the G/G/1. The results obtained in Chapter 4 are applied in Chapter 5 to analyse two-server tandem and cyclic queues. In Chapter 6 a maximum entropy model for general networks is built and implemented using the results of Chapter 4. In Chapter 7 it is illustrated how the maximum entropy analysis can improve the existing aggregation methods. Conclusions and comments about future prospects are given in Chapter 8.
Maximum entropy (ME) is a probability inference method that has been introduced in statistical mechanics by Jaynes [JAYN 57]. It is generally based on the entropy functional defined earlier by Shannon [SHAN 48] in communication theory. Since the analysis in the following chapters is strongly based on the maximum entropy formalism, this chapter will be devoted to the study of the method and the developments that led to its derivation. In section (2.1) a general historical survey is given. The properties of the entropy functional and its generalisation are discussed in section (2.2). The formalism of maximum entropy is presented in section (2.3). The discussion will be in the context of discrete systems with a brief note on applications to continuous systems.

2.1 An Overview

The concept of entropy was first introduced in thermodynamics more than a century ago [FAST 70]. Since then, it has been used in association with transformations from work effects to heat effects in that field. The thermodynamic or 'macroscopic' entropy is defined as

\[ S = \int_{0}^{T} \frac{dQ}{T} \]

where \( dQ \) is the increment of energy added to a body as heat during a
reversible process and \( T \) is the absolute temperature of the body during the reversible heat addition.

With advances in thermodynamics and the appearance of statistical mechanics, scientists became interested in the microstates corresponding to a macrostate of a thermodynamic system. Since the microstates are unmeasurable in general, people resorted to probability theory in analysing these systems. In this context the microscopic entropy was derived. The first expression for microscopic entropy appeared in the work of Boltzmann (1877) while he was looking for the most probable distribution of molecules in a conservative force field. Boltzmann considered a sample of gas consisting of a very large number, \( N \), of molecules distributed among \( c \) equal cells, the occupancy numbers being \( n_1, n_2, \ldots, n_c \) where \( n_i > 1 \) and \( \sum n_i = N \). He then defined the quantity

\[
H = k \ln W
\]  

(2.1)

where \( W \) is known as the disorder number and is defined as

\[
W = \frac{N!}{\prod_{i=1}^{c} n_i!}
\]

and \( k \) is the Boltzmann constant. Applying Stirling's approximation for the factorials, (2.1) becomes:

\[
H = -k N \sum_{i=1}^{c} \frac{n_i}{N} \ln \left( \frac{n_i}{N} \right)
\]  

(2.2)

It was soon realised that (2.2) has similar properties to the macroscopic entropy and therefore it became known as the microscopic entropy. Later Gibbs arrived at the same expression using an 'ensemble' of samples...
rather than one sample of the gas, and he suggested that (2.1) should be replaced by:

\[ H = k \ln \left( \frac{W}{N!} \right) \]

to avoid what is known as Gibbs's paradox.

Boltzmann was the first to observe that the entropy \( H \) is directly proportional to the disorder in the system. He even went further to maximise (2.2) in order to obtain the most probable distribution of molecules.

However it was the work of Shannon [SHAN 48] that eventually led to the use of entropy maximisation as a probability inference method. In his quest for a suitable uncertainty measure, Shannon considered a set of possible events whose probabilities of occurrence are \( p_1, p_2, \ldots, p_n \). He then demonstrated that the quantity which is positive, increases with uncertainty and is additive for independent sources of uncertainty:

\[ H(p) = -K \sum_i p_i \ln p_i \]  \hspace{1cm} (2.3)

where \( K \) is a positive constant. Because of the similarities between (2.2) and (2.3) Shannon called his function entropy.

Jaynes [JAYN 57] observed that Shannon's theory can be related to the principle of insufficient reason used earlier in probability theory by Bernoulli, Laplace and Bayes. The principle of insufficient reason states that: (i) a probability assignment is a state of knowledge, (ii) in determining a probability distribution, the outcomes of an event should be considered initially equally probable unless there is
evidence to let us think otherwise. Since Shannon's entropy attains its maximum when the outcomes of the event are equally likely, the principle of insufficient reason implies that in determining a probability distribution, one should start with the distribution of maximum entropy and then adjust this distribution in accordance with the information that becomes available, shifting the entropy maximum in the process. Accordingly, Jaynes concluded that: "given the propositions of an event and any information relating to them, the best estimate for the corresponding probabilities is the distribution that maximises the entropy subject to the available information", which is the principle of maximum entropy. Jaynes then described a formalism by which this distribution can be determined and demonstrated that this distribution is the least biased since it avoids any unjustifiable assumptions.

A few years later, Kullback [KULL 59] showed that maximum entropy can be generalised to include the case where there is an initial prior distribution \( g_i \) estimating the probabilities \( p_i \). One then maximises the quantity:

\[
H(p, g) = - \sum_i p_i \ln \frac{p_i}{g_i}
\]  

The expression (2.4) reduces to (2.3) when \( g_i \) is the uniform distribution [EVAN 79]. The quantity \(-H(p, g)\) is called cross-entropy or divergence.

For continuous systems it has been shown that the analog of (2.3) written as \(- \int p(x) \ln p(x) \, dx\) cannot apply because it fails to attain a finite maximum in certain situations [JAYN 68], [KOOP 79]. Jaynes [JAYN 68, 79]
showed that the proper expression to maximise is of the form

\[ H_x(p,m) = - \int p(x) \ln \left[ \frac{p(x)}{m(x)} \right] dx \]  

which is the analog of (2.4). However, in this case \( m(x) \) is an "invariant measure" that may not be a proper probability distribution and may not be known in advance. Up to now very little is known about the invariant measure \( m(x) \) and research is still going to investigate how it can be determined.

Note that maximisation of (2.4) is now used to obtain continuous probability distributions, but it clearly fails for symmetric (uniform prior) infinite continuous systems.

The principle of maximum entropy proved to be of great credibility in different fields of science. It is now successfully applied in statistical mechanics, information theory, spectral analysis, etc. For more on these applications, see the references in [SHOR 81]. Applications to queueing theory started as early as 1970. Ferdinand [FERD 70] applied maximum entropy to obtain the solution of the M/M/1//N system using analogy with statistical mechanics. Shore [SHOR 78] built an abstract model from which he obtained the maximum entropy solution for the M/M/∞ and M/M/∞//N queueing systems.

### 2.2 Basic Properties of the Entropy Functionals

Consider a system \( Q \) that has a set of possible discrete states \( S = \{ S_n \} \) where \( n \) may be finite or countably infinite. Let \( p(S_n) \) be the probability that the system \( Q \) is in state \( S_n \). The system entropy can then be defined as:
\[ H(p) = - \sum_{S_n \in S} p(S_n) \ln p(S_n) \]  

(2.6)

Clearly, \( H(p) \) has a minimum of zero when one of the probabilities \( \{p(S_n)\} \) is equal to unity, the others being zero i.e. the entropy is minimum when the uncertainty is minimum. On the other hand, \( H(p) \) is maximum when all the \( p(S_n) \) are equal i.e. when the uncertainty is maximum [SHAN 48]. This maximum exists if the number of possible states \( \{S_n\} \) is finite, in which case it is given by:

\[ H_{\text{max}}(p) = \ln N \]  

(2.7)

where \( N \) is the number of possible states. However, if the number of states is infinite, then this maximum does not exist unless other constraints are applied [KOOP 79] in the maximisation of \( H(p) \). Moreover, any change towards the equalisation of the \( p(S_n) \) increases the entropy \( H(p) \) i.e. the entropy increases with uncertainty [SHAN 48].

The above properties show that the entropy as defined by (2.6) is a measure of the distance between the distribution \( p(S_n) \) and the uniform distribution [JOHN 79] i.e. (2.6) is the entropy relative to the uniform distribution. This is true when the system considered is fundamentally symmetrical i.e. the states \( \{S_n\} \) are initially equally likely. An example of a fundamentally symmetrical system is the toss of an honest die where the probability of any face turning up is 1/6 [EVAN 79].

If initially there is a set of prior variables \( \{g(S_n)\} \) estimating the probabilities \( \{p(S_n)\} \), then the system is fundamentally non-symmetrical. An example of a fundamentally non-symmetrical system is the toss of a dishonest die in which 'two' is painted on four of the faces and 'three'
on the remaining two. For such systems, the entropy relative to the set of prior variable \( g(S_n) \) can be defined as

\[
H(p, g) = - \sum_{S_n \in S} p(S_n) \ln \left( \frac{p(S_n)}{g(S_n)} \right)
\]

which is the measure of the distance between the distribution \( p(S_n) \) and the prior variables \( g(S_n) \). If \( \sum_{S_n \in S} g(S_n) < \infty \) then the following theorem can be proved:

**Theorem (2.1)**

The relative entropy functional (2.8) is non-positive for any choice of \( p(S_n) \) and \( g(S_n) \) and has a maximum of zero at \( p(S_n) = g(S_n) \) for all \( S_n \).

**Proof**

Define \( \alpha(S_n) = g(S_n)/G \) where \( G \) is a normalising constant such that \( \sum_{S_n \in S} \alpha(S_n) = 1 \). Clearly \( g(S_n) = G\alpha(S_n) \) and \( \ln(g(S_n)) = \ln(G) + \ln(\alpha(S_n)) \).

Define \( L(S_n) = -p(S_n)\ln(p(S_n)) + p(S_n)\ln(G) + p(S_n)\ln(\alpha(S_n)) + p(S_n) - \alpha(S_n) \).

Clearly \( H(p, g) = \sum_{S_n \in S} L(S_n) \).

Differentiating \( L(S_n) \) with respect to \( p(S_n) \) we get

\[
\frac{\partial L(S_n)}{\partial p(S_n)} = -1 - \ln(p(S_n)) + \ln(G) + \ln(\alpha(S_n)) + 1
\]

\[
= \ln(g(S_n)) - \ln(p(S_n))
\]

Differentiating once more

\[
\frac{\partial^2 L(S_n)}{\partial p(S_n)^2} = - \frac{1}{p(S_n)}
\]
Since \( p(S_n) \) is always positive, this indicates that \( H(p,g) \) has a maximum of zero at \( p(S_n) = g(S_n) \), which also implies that \( H(p,g) \) is always non-positive.

Q.E.D.

It can easily be verified that the relative entropy functional (2.8) reduces to the entropy functional (2.6) when all the \( g(S_n) \) are equal.

In the particular case where the prior variables are of the form:

\[
g(S_n) = \begin{cases} 
1 & \text{if } S_n \in S_a \\
g_K = \text{constant} & \text{if } S_n \in S_b
\end{cases}
\]  

(2.9)

where \( \{S_a\} \) and \( \{S_b\} \) are two mutually exclusive and exhaustive partitions of \( \{S\} \); (2.8) can be expressed as

\[
H(p,g) = - \sum_{S_n \in S} p(S_n) \ln p(S_n) + \ln(g_K)(1 - \sum_{S_n \in S_a} p(S_n)) 
\]  

(2.10)

which will be monotone increasing as a function of \( H(p) \) if the sum

\[
\sum_{S_n \in S_a} p(S_n)
\]  

is fixed.

Note that if \( g(S_n) \) is a proper probability distribution then some authors consider the equivalent expression of 'cross-entropy' given by \((-H(p,g))\).

2.3 The Maximum Entropy Formalism

Consider the system \( Q \) defined in the above section. For this system, suppose that all what is known about the state probabilities \( p(S_n) \) are
(m+1) constraints of the form:

\[ \sum_{S_n \in S} p(S_n) = 1 \]  

(2.11)

and

\[ \sum_{S_n \in S} f_k(S_n) p(S_n) = F_k \quad 1 < k < m < \infty \]  

(2.12)

where \( \{F_k\} \) are expectations defined on a set of suitable functions \( \{f_k(S_n)\} \).

Since in general the number of constraints is less than the number of possible states, one is faced with an infinite number of distributions \( \{p(S_n)\} \) that satisfy these constraints. The problem is which one to choose.

The principle of maximum entropy states that: of all the distributions satisfying the constraints supplied by the given information, the minimally prejudiced distribution which should be chosen is the one that maximises the system entropy (2.6) subject to the constraints (2.11) and (2.12).

Maximisation of (2.6) applies only when the system is fundamentally symmetrical, otherwise one should maximise the relative entropy (2.8) subject to the constraints (2.11) and (2.12).

The maximisation is usually carried out using Lagrange’s method of undetermined multipliers. Defining \( \beta_0 \) to be the Lagrange multiplier corresponding to the normalisation constraint (2.11) and \( \{\beta_k\} \) to be the set of Lagrange multipliers corresponding to the set of constraints (2.12), then the Lagrangian of (2.8), (2.11) and (2.12) can be formed as follows (\text{TRIB 69} p.123):

\[ \sum_{S_n \in S} \ln p(S_n) - \ln g(S_n) + \beta_0 + \sum_{k=1}^{m+1} \beta_k f_k(S_n) = 0 \]

or

\[ \ln(p(S_n)) - \ln (g(S_n)) + \beta_0 + \sum_k \beta_k f_k(S_n) = 0 \]
which gives

$$\ln p(S_n) = \ln g(S_n) - \beta_0 - \sum_k \beta_k f_k(S_n)$$

or

$$p(S_n) = g(S_n) \exp(-\beta_0 - \sum_k \beta_k f_k(S_n))$$

(2.13)

where $Z_p$, known in statistical mechanics as the partition function, is given by

$$Z_p = \exp(\beta_0) = \sum_{S_n \in S} \exp(-\sum_k \beta_k f_k(S_n))$$

(2.14)

The Lagrange multipliers $\{\beta_k\}$ are often determined using the relation

$$\frac{\partial \ln(Z_p)}{\partial \beta_k} = F_k$$

(2.15)

which is another way of expressing the constraints (2.11).

Of course, the values of the prior variables $g(S_n)$ may not all be known apriori, but it may be known that these variables exist. Based on prior knowledge about the states of the system (in addition to the constraints), either the value or the type of each $g(S_n)$ may be established. This information therefore can be incorporated into the maximum entropy formalism to determine the form of the probability distribution $\{p(S_n)\}$ expressed by (2.13) and (2.14). However, this will not determine the numerical values of those still unknown $\{g(S_n)\}$ unless they can be computed as a result of the optimisation process. Furthermore, we may not always know the expectations $\{F_k\}$ associated with the constraints (2.12), but we may know
that such expectations exist. This information may similarly be fed into the formalism to contribute towards determining the form of the probability distribution, but the numerical values of the constants appearing in the solution will have to be supplied at a later stage ([TRIB 69], p.122).

For continuous systems (2.8) generalises correctly, and one can obtain the required probability distribution by maximising:

$$H_x(p,g) = - \int_x p(x) \ln \left[ \frac{p(x)}{g(x)} \right]$$

subject to the constraints

$$\int_x p(x) \, dx = 1$$

and

$$\int_x f_k(x) \, p(x) = F_k$$

where $g(x)$ here is the "invariant measure" defined by Jaynes [JAYN 68]. Although this invariant measure is known to exist, very little is known about its nature or how it can be determined. Research is still going on [JAYN 79] to help determine these variables (using the method of marginalisation and transformation groups).

For more on maximum entropy and its generalisations, see the references cited above.
CHAPTER III
MAXIMUM ENTROPY ANALYSIS OF THE M/G/l AND THE G/M/l QUEUEING SYSTEMS IN EQUILIBRIUM

3.1 Introduction

The M/G/1 queueing model represents an infinite capacity queueing system with random (Poisson) arrivals and a general (G-type) service time distribution. The G/M/1 queueing model is the dual of the M/G/1 with a G-type arrival pattern and a single exponential server. These models are of great value in the performance analysis of two-server cyclic queueing models of multiprogramming |REIS 74| |ALLE 78| and the aggregation of general queueing networks |COUR 77|. For both models the equilibrium solution for the number of jobs in the system varies with the probability distribution function chosen to represent the G-type distribution. Even in the presence of empirical data, the characterisation of this function involves a degree of arbitrariness that may cause some variation in the results |LAZO 77|. For example, there is an infinite number of two-stage models representing a G-type distribution with the same mean value and coefficient of variation. Analysts in practice choose one of these models by making an arbitrary parameter selection, typically for algebraic convenience |SAUE 81|.

In this chapter, the maximum entropy formalism is used in order to provide a solution to this problem. A generalised maximum entropy model is directly applied to derive a unique solution for the M/G/1 - and G/M/1 - queueing systems at equilibrium. The motivation is to show how the maximum entropy formalism can relieve the analyst from making
The maximum entropy solution for the $M/G/1$ queueing system and the corresponding service time distribution are obtained in section 3.2. The maximum entropy solution for the $G/M/1$ is obtained in section 3.3. The balance equations associated with these solutions are produced in section 3.4. The relation between the maximum entropy solution and the solutions obtained using the method of stages is discussed in section 3.5. A concluding summary is given in section 3.6.

3.2 Maximum Entropy Solution for the $M/G/1$ Queueing System

In this section a generalised maximum entropy model is used to derive a unique solution for the $M/G/1$ queueing system.

The constraints of the system are first defined in section 3.2.1. These constraints are then used in section 3.2.2 to obtain the maximum entropy solution. The underlying service time distribution is obtained in section 3.2.3.

3.2.1 The constraints of the $M/G/1$

There are two basic results from classical queueing theory that facilitate the application of maximum entropy formalism to the analysis of the $M/G/1$ queueing system.

The first result is the probability of having an empty system given as

$$p_0 = 1 - \rho$$

where $\rho$ is the utilisation of the system.
The second result is the Pollaczek-Khinchin formula for the mean number of jobs in the system given as

\[ <n> = \rho + \frac{\rho^2 (1 + C_s^2)}{2(1 - \rho)} \]  
\hspace{1cm} (3.2)

where \( C_s^2 \) is the service time squared coefficient of variation. Detailed derivations of (3.1) and (3.2) can be found in [KLEI 76a][ALLE 78].

Defining \( y_s \) as

\[ y_s = (C_s^2 - 1)/2 \]  
\hspace{1cm} (3.3)

expression (3.2) can be re-written as

\[ <n> = \frac{\rho(1 + \rho y_s)}{(1 - \rho)} \]  
\hspace{1cm} (3.4)

As mentioned earlier, there is an infinite number of distributions that satisfy both (3.1) and (3.4) for the same value of \( \rho \) and \( y_s \), and can therefore be considered as solutions for the M/G/1 model. According to the maximum entropy formalism, the one that maximises the system's entropy should be chosen.

3.2.2 The maximum entropy model

Following the above discussion, the maximum entropy solution for the M/G/1 can be obtained by maximising (2.6) written as:

\[ H(p) = - \sum_{n=0}^{\infty} p_n \ln p_n \]  
\hspace{1cm} (3.5)

subject to the constraints
Given this information, the following theorem can be presented:

**Theorem 3.1**

The maximum entropy solution for the M/G/1 is given by:

\[
        p_n = \begin{cases} 
        1 - \rho & n=0 \\
        (1 - \rho)gx^n & n>0 
        \end{cases} \tag{3.9}
\]

where

\[
x = \frac{\rho(1 + y_s)}{1 + \rho y_s} \tag{3.10}
\]

and

\[
g = 1/(1+y_s) \tag{3.11}
\]

Proof

Maximising (3.5) subject to the constraints (3.6)-(3.8) given by (2.13):

\[
p_n = \frac{1}{Z_p} \exp \{ -\beta_1 K(n) - \beta_2 n \} \tag{3.12}
\]

where, by (2.14)

\[
Z_p = \sum_{n=0}^{\infty} \exp \{ -\beta_1 K(n) - \beta_2 n \} \tag{3.13}
\]
and $\beta_1$, $\beta_2$ are the Lagrange multipliers corresponding to the constraints (3.7), (3.8) respectively. For $n=0$, (3.12) gives

$$Z_p^{-1} = p_0 = (1-p)$$

(3.14)

Defining $g = \exp(-\beta_1)$, $x = \exp(-\beta_2)$ and using (3.12), (3.14) in (3.7), (3.8) we obtain the following two equations:

$$\frac{g^x}{1-x} = \frac{\rho}{1-p}$$

$$\frac{g^x}{(1-x)^2} = \frac{\rho(1+\rho y_s)}{(1-p)^2}$$

Solving for $g$ and $x$ we get

$$g = 1/(1+y_s) \quad , \quad x = \rho(1+y_s)/(1+\rho y_s)$$

Q.E.D

More details can be found in [ELAF 82] and [ELAF 83a]

As an example, consider the M/M/1 queueing system. Since $C_s^2 = 1$, it is implied that:

$$y_s = 0, x = \rho, \beta = -\ln \{\rho\} \quad \text{and} \quad r_s = 1.$$ 

Therefore, from (3.9) $k_n = 1$, for all $n$, and (3.9) gives

$$p_n = (1-\rho)^n$$

which is the classical result. Note that the M/M/1 is, as the intuition suggests, the only fundamentally symmetrical M/G/1 system with a prior variable $g_n = 1$ for all $n = 0, 1, 2, \ldots$.

3.2.3 The underlying service time distribution: A GE distribution

The unique solution (3.9) has been derived without mentioning the pdf of the service time distribution. The immediate question that arises is what is the role of this function in the solution (3.9). The answer to this question is given by the following investigation, where the pdf of the service time distribution is determined.
Theorem 3.2

The maximum entropy M/G/l solution (3.9) is equivalent to the equilibrium solution of an M/G/l queueing system with a service time distribution of the form:

\[ f(t) = (1 - r_s) u(t) + r_s \mu' \exp \{-\mu' t\} \] (3.15)

where

\[ \mu' = r_s \mu, \quad r_s = 1/(1 + y_s) \] (3.16)

and \( u(t) \) is the unit impulse function, defined by

\[ u(t) = \begin{cases} 1, & t=0 \\ 0, & t \neq 0 \end{cases} \]

such that \( \int_{-\infty}^{\infty} u(t) dt = 1 \).

Proof

It is known that the \( \varphi \)-transform of the equilibrium solution of any M/G/l queueing system is given by the Pollaczek–Khinchin transform equation:

\[ Q(z) = \frac{F^*(\lambda - \lambda z)(1-p)(1-z)}{F^*(\lambda - \lambda z) - z} \] (3.17)

where \( F^*(\theta) \) is the Laplace transform of the service time distribution.

For the solution (3.9) this transform can be derived directly using the relation

\[ * \text{Note that for } C^2 < 1, \text{ the function } f(t) \text{ includes a negative zero probability } ((1-r_s) u(t)) \text{ which makes it an improper distribution. However the ME distribution } p_n \text{ is unaffected.} \]
\[ Q(z) = \sum_{n=0}^{\infty} p_n z^n, \quad |z| < 1 \]

This implies that

\[
Q(z) = (1-p) + \frac{(1-p)xz}{(y_s+1)(1-xz)}
= \frac{(1-p)(1-xz(1-r_s))}{1-xz}
\]

(3.18)

where \(x\) and \(r_s\) are given by (3.10) and (3.16) respectively. It can be easily verified that \(Q(0) = 1-p\) and \(Q(1) = 1\). Equating the right-hand sides of (3.17) and (3.18), substituting for \(x\) and solving for \(F^*(\lambda-\lambda z)\), the following result is obtained:

\[
F^*(\lambda-\lambda z) = \frac{r_s \mu + (1-r_s)(\lambda-\lambda z)}{r_s \mu + \lambda-\lambda z}
\]

Substituting \(\theta\) for \((\lambda-\lambda z)\), the above relation becomes

\[
F^*(\theta) = \frac{r_s \mu + (1-r_s)\theta}{r_s \mu + \theta}
= (1-r_s) + \frac{\mu}{r_s \mu + \theta}
\]

Inverting, result (3.15) follows.

Q.E.D.
The following parameters for the distribution of the service time $S$ are easily obtained:

The mean of $S$:

$$E(S) = \int_0^\infty t f(t) \, dt = \frac{1}{\mu}$$

The second moment of $S$:

$$E(S^2) = \int_0^\infty t^2 f(t) \, dt = \frac{2}{r_s^2 \mu^2}$$

The squared coefficient of variation of $S$:

$$c_s^2 = \frac{E(S^2)}{[E(S)]^2} - 1 = \frac{2-r_s}{r_s}$$

The pdf is

$$F(t) = \int_{-\infty}^{t} f(t) = 1 - r_s \exp(-\mu t) \quad (3.19)$$

Throughout this study the distribution (3.15) will be called the generalised exponential or shortly the GE distribution. This is because it involves a generalised function (unit impulse function) at the origin and because it spans a wide family of exponential distribution.
3.3 Maximum Entropy Formalism and the G/M/1

3.3.1 Statement of the problem

The equilibrium probability of finding \( n \) jobs in a G/M/1 queueing system \cite{klei75} is given by:

\[
p_n = \begin{cases} 
1 - \rho & \text{n=0} \\
\frac{\rho}{1-\sigma} \sigma^{n-1} & \text{n>0}
\end{cases}
\]

(3.20)

where \( \rho \) is the server utilisation and \( \sigma \) satisfies

\[
\sigma = A^* (\mu - \mu \sigma), \quad 0 < \sigma < 1
\]

(3.21)

and can be interpreted as the probability that an arriving job finds a busy server; \( A^*(\theta) \) is the Laplace transform of the inter-arrival time distribution and \( \mu \) is the mean service rate. Furthermore, the mean number of jobs in the system is given by

\[
<n> = \sum_{n=0}^{\infty} n p_n = \frac{\rho}{1-\sigma}
\]

(3.22)

In common practice, the relation (3.21) is specified by selecting a two-stage representation for the inter-arrival time distribution. The problem which arises is similar to that of the M/G/1 model since there are infinitely many pdf's satisfying the constraints. Hence the value of \( \sigma \) and consequently the distribution (3.20) is not unique, causing variation in the performance metrics.
3.3.2 A maximum entropy solution

Following the discussion for the maximum entropy M/G/1 it can be conjectured that an inter-arrival time distribution of the type (3.15) produces a maximum entropy solution for the G/M/1 system in equilibrium. Later in the chapter it is shown that this conjecture is substantially true. Based on this, a unique value of \( \sigma \) can be determined as follows.

Theorem 3.4

The value of \( \sigma \) that corresponds to the maximum entropy inter-arrival time distribution of the type

\[
a(t) = (1-r_a)u_0(t) + r_a \lambda' \exp \{-\lambda't\} \]

(3.23)

\[
\lambda' = r_a \lambda
\]

is given by

\[
\sigma = 1 - r_a (1-p)
\]

(3.24)

where

\[
r_a = 1/(1 + y_a)
\]

(3.25)

\[
y_a = (C_a^2 -1)/2
\]

(3.26)

\( C_a \) is the inter-arrival time coefficient of variation and \( u_0(t) \) is the unit impulse function.

Proof

The Laplace transform of the distribution (3.23) is given by

\[
A^*(\theta) = (1-r_a) + \frac{r_a \lambda'}{\lambda' + \theta}
\]

(3.27)

Note that the constraint \( 0 < \sigma < 1 \) will mean that \( 0 < 1-r_a (1-p) < 1 \) which implies that the solution \( \rho \) applies for \( p \) in the range: \( (1-C_a^2)/2 < p < 1 \).
Hence (3.21) becomes
\[ \sigma = (1-r_a) + \frac{r_a \lambda'}{\lambda' + \mu(1-\sigma)} \]
which reduces to
\[ \mu \sigma = \lambda' + \mu - r_a \mu \]
Solving for \( \sigma \) yields result (3.24).

\[ \text{Q.E.D.} \]

The maximum entropy solution for the G/M/1 queueing system is now formed by maximising \( H(p) \) given by (3.5) subject to the constraints (3.6), (3.7) and (3.22), which by (3.24), becomes:
\[ \sum_{n=0}^{\infty} n p_n = \frac{\rho}{r_a} (1-\rho) \quad (3.28) \]
This solution can be obtained by the following theorem.

**Theorem 3.5**

The maximum entropy solution for the G/M/1 is given by
\[ p_n = \begin{cases} 
(1 - \rho) & n = 0 \\
(1 - \rho)g \times \frac{n}{n} & n > 0 
\end{cases} \quad (3.29) \]
where

\[ x = \frac{\rho + y_a}{1 + y_a} \tag{3.30} \]

and

\[ g = \frac{\rho}{\rho + y_a} \tag{3.31} \]

**Proof**

Making the same argument as for the M/G/1, it can be established that the G/M/l maximum entropy solution is of the form

\[ p_n = \begin{cases} (1-p) & n=0 \\ (1-p)g^n & n>0 \end{cases} \]

Taking this into the normalisation and mean value constraint (3.28) we get

\[ \frac{gx}{1-x} = \frac{\rho}{1-p} \]

\[ \frac{g}{(1-x)^2} = \frac{\rho}{r_a (1-p)} = \frac{\rho (1+y_a)}{1-p} \]

Solving, we get

\[ x = \frac{(\rho + y_a)}{(1+y_a)} \]

and

\[ g = \frac{\rho}{(y_a + \rho)} \]

and the theorem follows.

Q.E.D.

**Example**

In Table 1 below, the average number of jobs in an E_2/M/l is computed, using the maximum entropy result <n> given by (3.28), the
Table 3.1 The average number of customers in the GE/M/I compared to that of the E₂/M/I and the diffusion approximation

<table>
<thead>
<tr>
<th>ρ</th>
<th>&lt;n&gt; = \frac{ρ(1+y_a)}{1-ρ}</th>
<th>&lt;n&gt;_E = \frac{ρ}{-2ρ+0.5+y_a/2ρ+0.25}</th>
<th>&lt;n&gt;_D = \frac{ρ(1+py_a)}{1-ρ}</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.95</td>
<td>14.2500</td>
<td>14.3314</td>
<td>14.4875</td>
</tr>
<tr>
<td>0.90</td>
<td>6.7500</td>
<td>6.8295</td>
<td>6.9750</td>
</tr>
<tr>
<td>0.85</td>
<td>4.2500</td>
<td>4.3274</td>
<td>4.4625</td>
</tr>
<tr>
<td>0.80</td>
<td>3.0000</td>
<td>3.0752</td>
<td>3.2000</td>
</tr>
<tr>
<td>0.75</td>
<td>2.2500</td>
<td>2.3229</td>
<td>2.4375</td>
</tr>
<tr>
<td>0.70</td>
<td>1.7500</td>
<td>1.8204</td>
<td>1.9250</td>
</tr>
<tr>
<td>0.60</td>
<td>1.1250</td>
<td>1.1901</td>
<td>1.2750</td>
</tr>
<tr>
<td>0.50</td>
<td>0.7500</td>
<td>0.8090</td>
<td>0.8750</td>
</tr>
<tr>
<td>0.40</td>
<td>0.5000</td>
<td>0.5520</td>
<td>0.6000</td>
</tr>
<tr>
<td>0.30</td>
<td>0.3214</td>
<td>0.3650</td>
<td>0.3964</td>
</tr>
<tr>
<td>0.20</td>
<td>0.1875 *</td>
<td>0.2207</td>
<td>0.2375</td>
</tr>
<tr>
<td>0.10</td>
<td>0.0833 *</td>
<td>0.1030</td>
<td>0.1083</td>
</tr>
</tbody>
</table>

*These values are improper since they will imply that <n> < ρ which contradicts basic queueing theory results. This is because the condition ρ > (1 - C)^{1/2}.

31.
method of stages result $<n>_E$ and the diffusion approximation $<n>_D$

The results are compared for different values of $\rho$. For the method of stages an Erlang-2 distribution is used.

It can be verified that

$$<n>_D - <n> = -y_a \rho$$

which means that the deviation of the diffusion approximation from the maximum entropy results is a function of $\rho$. Note that in both methods the percentage deviation from maximum entropy grows as $\rho$ decreases.

The difference may be more dramatic in the case of the G/M/1/$N$. The percentage deviation in the diffusion approximation ranges from 1.66% for $\rho = 0.95$ to 30.48% for $\rho = 0.1$. For the phase method (Erlang type) the percentage deviation from maximum entropy ranges from 0.56% for $\rho = 0.95$ to 24.1% for $\rho = 0.1$. For $\rho = 0.7$ the percentage deviation from maximum entropy is 10% for the diffusion approximation and 4% for the phase method.

### 3.4 Balance Equations for the Maximum Entropy M/G/1 - and G/M/1 - Systems

The equilibrium maximum entropy solution for the M/G/1 - or G/M/1 - queueing systems, given by (3.9) and (3.29) respectively, can be uniformly represented by

$$p_n = c_n \psi_n^n$$

where

$$c_n = (1-\rho)(\gamma_o/\gamma_n)$$

and

$$\psi_n = \lambda_n/\mu_n$$
### Table 3.2
Parameters of Maximum Entropy M/G/1 - and G/M/1 - Systems

<table>
<thead>
<tr>
<th>Parameters</th>
<th>M/G/1</th>
<th>G/M/1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma_0$</td>
<td>$r_s$</td>
<td>$\rho/(\gamma_a+\rho)$</td>
</tr>
<tr>
<td>$\gamma_n, n \geq 0$</td>
<td>$1$</td>
<td>$1$</td>
</tr>
<tr>
<td>$\lambda_0$</td>
<td>$r_s \lambda$</td>
<td>$r_a \lambda$</td>
</tr>
<tr>
<td>$\lambda_g = \lambda_n, n \geq 0$</td>
<td>$\lambda$</td>
<td>$r_a \lambda + (1-r_a) \mu$</td>
</tr>
<tr>
<td>$\mu_g = \mu_n, all n$</td>
<td>$r_s \mu + (1-r_s) \lambda$</td>
<td>$\mu$</td>
</tr>
</tbody>
</table>

Where $\rho = \lambda/\mu$, $r_s = 1/(1+y_s)$, $r_a = 1/(1+y_a)$, $y_s = (c_s^2 - 1)/2$, $y_a = (c_a^2 - 1)/2$.
Fig. 3.1(a) Balance diagram for M/G/1

Fig. 3.1(b) Balance diagram for G/M/1
where $y_n$, $\lambda_n$ and $\mu_n$ are described in Table 3.2.

From (3.32)-(3.34) and Table 3.2, the balance diagrams shown in Fig. 3.1 can be drawn. From these diagrams, the following set of balance equations can be obtained:

\[
\begin{align*}
\lambda_0 p_0 &= \mu_g p_1 \\
\lambda_g p_{n-1} + \mu_g p_{n+1} &= (\lambda + \mu_g) p_n & (n > 0)
\end{align*}
\]

From (3.35), it is clear that the maximum entropy $M/G/1$ - and $G/M/1$ - queueing systems satisfy local balance.

### 3.5 Maximum Entropy and the Method of Stages

As mentioned in the Introduction, the equilibrium solution to the $M/G/1$ - $G/M/1$ - queueing system is often obtained using a suitable stage-representation for the service - arrival - time distribution. There are two well-known models in this respect: the hyperexponential-2 which is used when the squared coefficient of variation is greater than one, and the hypoexponential-2 which is chosen when the squared coefficient of variation is less than one. These two models are not completely disjoint and in fact belong to the same family if some parameters of the hypoexponential-2 are allowed to take complex values [KLEI 75 | SAUE 81].

In this section the relation between the maximum entropy solutions obtained in the previous sections and the solutions obtained using the method of stages is investigated. It is shown that the maximum entropy of any two-stage $M/G/1$ converges to solution (3.9) when the system parameters tend to the limits specified below. Furthermore, it is proved that for the $G/M/1$, any solution produced by a two-stage inter-
arrival time distribution has an entropy lower than that of solution (3.29). In this sense, the conjecture of section 3.3.1 holds.

3.5.1 Parameters of the two-stage systems

Maintaining the generalisation mentioned above, the service or inter-arrival time distribution of any two-stage system can be represented by the following pdf:

\[ h(t) = a_1 v_1 \exp\{-v_1 t\} + a_2 v_2 \exp\{-v_2 t\} \]  \hspace{1cm} (3.36)

where \( v_1 \) and \( v_2 \) (the mean service or arrival rates of the two imaginary servers) satisfy the relation:

\[ \left(\frac{a_1}{v_1}\right) + \left(\frac{a_2}{v_2}\right) = \frac{1}{\nu} \]  \hspace{1cm} (3.37)

and

\[ C^2 = \frac{2\left\{\frac{a_1}{v_1} + \frac{a_2}{v_2}\right\}}{\left\{\frac{a_1}{v_1} + \frac{a_2}{v_2}\right\}^2} - 1 \]  \hspace{1cm} (3.38)

\( \nu \) being the mean service or arrival rate, \( C^2 \) is the squared coefficient of variation, and

\[ a_1 + a_2 = 1 \]  \hspace{1cm} (3.39)

If \( C^2 > 1 \), then \( 0 < a_1, a_2 < 1 \). However, if \( C^2 < 1 \), then \( a_1, a_2 \) have opposite signs with the additional constraint
\[
\alpha_1 = \frac{v_1}{v_2-v_1}, \quad \alpha_2 = \frac{v_2}{v_1-v_2}, \quad v_1 \neq v_2 \quad (3.40)
\]

Constraint (3.37) is satisfied for

\[
\begin{align*}
\nu_1 &= k\alpha_1 \nu \\
\nu_2 &= k\alpha_2 \nu / (k-1)
\end{align*}
\quad (3.41)
\]

where \(1 < k < \infty\) for \(C^2 > 1\) and \(-\infty < k < 1\) for \(C^2 < 1\) (because of (3.51)). Taking (3.41) into (3.38) and solving for \(\alpha_1\), the following expression is obtained:

\[
\alpha_1 = \frac{C^2 - 1}{2(C^2 + 1)} + \frac{2}{k(C^2 + 1)} + \frac{(C^2 - 1)^2 + 8(C^2 - 1)/k + 8(1-C^2)/k^2}{2(C^2 + 1)}
\quad (3.41)
\]

The following theorem can now be presented:

**Theorem 3.6**

For any two-stage M/G/I queueing system,

A. \(\lim_{|k|\to\infty} \alpha_1 = (C^2 - 1)/(C^2 + 1) = y/(y+1)\) \quad (3.42)

B. \(|\alpha_1| > |y|/(y+1), \quad \alpha_2 < 1/(y+1)\)

\[|\alpha_1|/\alpha_2 > |y|, \quad \alpha_2/|\alpha_1| > 1/|y|\] \quad (3.43)

where \(y = (C^2 - 1)/2\).

**Proof**

Taking the limit as \(|k| \to \infty\) the following two solutions are obtained.

\[\lim \alpha_1 = (C^2 - 1)/(C^2 + 1)\] or \(\alpha_1 = 0\)
The second root is rejected since the value of $C^2$ is not necessarily one.

Using (3.3), $\lim \alpha_1 = y/(y+1)$ and since $\alpha_1 + \alpha_2 = 1$, $\lim \alpha_2 = 1/(y+1)$. Taking the absolute values, it is clear that $|y|/(y+1)$ is always a lower limit for $|\alpha_1|$ and $1/(y+1)$ is always an upper limit for $\alpha_2$. Therefore, $|\alpha_1|/\alpha_2 > |y|$ and $\alpha_2/|\alpha_1| > 1/|y|$. 

Q.E.D.

3.5.2 Convergence of the two-stage M/G/1

For all two-stage M/G/1 systems satisfying the above assumptions, it has been shown that the equilibrium solution [ALLE 78] is of the "sum form":

$$q_n = (1-\rho)\alpha_1 (1/z_1)^n + (1-\rho)\alpha_2 (1/z_2)^n, \quad n = 0,1,2,...$$

where $0 < 1/z_1$, $1/z_2 < 1$.

Writing

$$x_1 = 1/z_1, \quad x_2 = 1/z_2$$

this solution can be rewritten as

$$q_n = (1-\rho)(\alpha_1 x_2^n + \alpha_2 x_2^n), \quad n = 0,1,2,...$$

(3.44)

Based on this "prior" information, it is shown below that the equilibrium maximum entropy solution of these systems converges to (3.16) as $\alpha_1 \to y_s/(y_s+1)$, where $y$ in Theorem 3.6 is now replaced by $y_s = (C^2-1)/2$. 
Theorem 3.7

A. The equilibrium maximum entropy solution to any two-stage M/G/1 system subject to the constraints of the form (3.6) and (3.8) is given by

\[ q_n = (1-\rho) \left\{ a_1 \left[ \frac{\rho (1-\sqrt{y_s a_2/a_1})}{1-\rho \sqrt{y_s a_2/a_1}} \right]^n + a_2 \left[ \frac{\rho (1+\sqrt{y_s a_2/a_1})}{1+\rho \sqrt{y_s a_2/a_1}} \right]^n \right\} \quad (3.45) \]

\[ n = 0, 1, 2, \ldots \]

B. \( \lim q_n = p_n \)

\[ a_1 \rightarrow y_s / (y_s + 1) \quad (3.46) \]

where \( p_n \) is the maximum entropy M/G/1 solution given by (3.9)-(3.11).

Proof

The maximum entropy solution (3.44) to any two-stage M/G/1 system can be obtained by maximising the entropy (3.5) written as

\[ H(q) = - \sum_{n=0}^{\infty} q_n \ln q_n \quad (3.47) \]

subject to the constraints (3.6) and (3.8) written as

\[ \sum_{n=0}^{\infty} q_n = 1 \]

\[ \sum_{n=0}^{\infty} n q_n = \frac{\rho (1+y_s)}{1-\rho} \quad (3.48) \]

Substituting (3.44) into (3.48) and (3.49) the following two relations are established:
\[
\frac{a_1}{1-x_1} + \frac{a_2}{1-x_2} = z_p
\]

(3.50)

and

\[
\frac{a_1 x_1}{(1-x_1)^2} + \frac{a_2 x_2}{(1-x_2)^2} = \frac{\rho (1 + \rho y_s) z_p}{1 - \rho}
\]

(3.51)

where \( z_p = (1-\rho)^{-1} \).

Using the change of variable \( S_1 = 1/(1-x_1) \), \( S_2 = 1/(1-x_2) \) and the substitution

\[
S_2 = \frac{(z_p - a_1 S_1)}{a_2}
\]

(3.52)

the following quadratic equation in \( S_1 \) is obtained:

\[
\frac{a_1}{a_2} S_1^2 - 2 \frac{a_2}{a_2^2} S_1 + \frac{z_p^2}{a_2} - z_p - \rho (1 + \rho y_s) z_p^2 = 0
\]

(3.53)

Solving for \( S_1 \), the roots are

\[
S_1 = z_p \left| 1 + \rho \left( \frac{y_s}{a_1} \right)^{1/2} \right|
\]

(3.54)

Substituting for \( x_1 \) and \( x_2 \), (3.44) produces:

\[
(1-\rho) \left\{ a_1 \left[ \frac{\rho \left(1 + y_s \frac{a_2}{a_1} \right)}{1 + \rho y_s \frac{a_2}{a_1}} \right]^n + a_2 \left[ \frac{\rho \left(1 - y_s \frac{a_2}{a_1} \right)}{1 - \rho y_s \frac{a_2}{a_1}} \right]^n \right\}
\]

if \( S_1 = z_p \left(1 + \rho y_s \frac{a_2}{a_1} \right) \)

\[
q_n = \{
\]

(1-\rho) \left\{ a_1 \left[ \frac{\rho \left(1 - y_s \frac{a_2}{a_1} \right)}{1 - \rho y_s \frac{a_2}{a_1}} \right]^n + a_2 \left[ \frac{\rho \left(1 + y_s \frac{a_2}{a_1} \right)}{1 + \rho y_s \frac{a_2}{a_1}} \right]^n \right\}
\]

if \( S_1 = z_p \left(1 - \rho y_s \frac{a_2}{a_1} \right) \)
Taking the limit as \( y_s \to 1+y_s \), (3.54) becomes

\[
(1-p)\left\{ \frac{y_s}{y_s+1} \left[ \frac{2p}{1+p} \right]^n + \frac{1}{y_s+1} \left[ \frac{p(1-y_s)}{1-y_s} \right]^n \right\}
\]

\( q_n = \frac{\rho(y_s+1)}{1+y_s} \)

where, here, \( k_n = \begin{cases} 1/(1+y_s) & n=0 \\ 1 & n>0 \end{cases} \)

The second solution is the maximum entropy solution (3.9) derived for the M/G/1 in section

Since there is only one maximum entropy solution [SHOR 81] [TRIB 69], the other solution is eliminated. This completes the proof.

Q.E.D.

3.5.3 Maximum entropy and the two-stage G/M/1

In this section it is shown that the equilibrium solution of any two-stage G/M/1 converges to the solution (3.29-3.31) and simultaneously the system entropy \( H(p) \) converges to its maximum. The analysis is meant to confirm the conjecture made in Section 3.3.2.

First, the following theorem is presented.

Theorem 3.8

For the G/M/1 model, the system entropy is monotone increasing as a function of \( \sigma \) in the interval \((\phi,1)\).

Proof

The system entropy is given by:

\[
H(p) = - \sum_{n=0}^{\infty} p_n \ln p_n
\]
Using (3.20) this can be transformed into:

$$H(\sigma) = -\ln (1-\rho) - \rho \ln\left(\frac{\rho}{1-\rho}\right) - \rho \ln (1-\sigma) - \frac{\rho \sigma \ln \sigma}{1-\sigma}$$

Differentiating w.r.t. $\sigma$ we get

$$\frac{\partial H(\sigma)}{\partial \sigma} = -\frac{\rho \ln \sigma}{(1-\sigma)^2}$$

Since this expression is $>0$ for $\sigma$ in the interval $(0,1)$, then by the mean value theorem of differential calculus \[COUR 65\] $H(\sigma)$ is monotone increasing as a function of $\sigma$.

**Corollary**

The maximum entropy solution for the G/M/I corresponds to the largest over all the values of $\sigma$ that satisfy (3.21) and (3.22).

To find this solution, an explicit expression for $\sigma$ is first sought. Such an expression can be formulated using (3.21), (3.36) and (3.41). The Laplace transform of the interarrival time distribution of any two-stage G/M/I can be obtained from (3.36) as:

$$A^x(\theta) = \frac{a_1 \lambda_1}{\lambda_1 + \theta} + \frac{a_2 \lambda_2}{\lambda_2 + \theta}$$

where $v_i$ is replaced by $\lambda_i$, $i = 1,2$.

Using (3.41) and replacing $v$ by $\lambda$ we get:
\[
A^\lambda(\theta) = \frac{\kappa a_1^2}{\kappa a_1^2 + \kappa a_2^2 \lambda + \theta} \quad \text{(3.56)}
\]

from which

\[
A^\lambda(\mu - \sigma) = \frac{\kappa a_1^2}{\kappa a_1^2 + \mu (1 - \sigma)} \quad \text{(3.57)}
\]

Using (3.3) and (3.15) we obtain the following cubic in \( \sigma \):

\[
(k-1)\sigma^3 - (k\rho (ka_1 + \alpha_2 - \alpha_1) + 2(k-1)\sigma^2
\]

\[
+ (2k\rho (\alpha_2 - \alpha_1) + k^2 \rho (\alpha_2 + \alpha_1) + (k-1))\sigma
\]

\[
- (k^2 m_1 m_2 + k\rho (k a_2^2 - \alpha_2 - \alpha_1)) = \phi
\]

It can be checked that \( \sigma = 1 \) is always a root of this polynomial. Therefore, we divide by \( (1-\sigma) \) to obtain the following quadratic equation in \( \sigma \):

\[
(k-1)\sigma^2 - (k\rho (ka_1 + \alpha_2 - \alpha_1) + (k-1))\sigma
\]

\[
+ (k\rho (\alpha_2 - \alpha_1) + k^2 \rho (\alpha_2 + \alpha_1)) = \phi
\]

Dividing all through by \( k^2 \) this becomes:

\[
\frac{(k-1)}{k^2} \sigma^2 - (\rho (\alpha_1 + k\frac{\alpha_2 - \alpha_1}{k}) + \frac{(k-1)}{k^2}) \sigma
\]

\[
+ \left( \frac{\rho}{k} (\alpha_2 - \alpha_1) + \alpha_1 \rho (\alpha_2 + \alpha_1) \right) = \phi
\]

Defining \( A = (k-1)/k^2 \), \( B = -\frac{\rho}{k} (ka_1 + \alpha_2 - \alpha_1) - \frac{(k-1)}{k^2} \),

\[
C = \frac{\rho}{k} (\alpha_2 - \alpha_1) + \alpha_1 \rho (\alpha_2 + \alpha_1)
\]

this can be rewritten as:
Theorem 3.9

For the G/M/1 queueing system: if \( \frac{C_a}{\alpha} \geq 1 \) then:

a. the upper bound for \( \sigma \) in the interval (0,1) is given by

\[
\sigma = 1 - r_a \left(1 - \rho \right)
\]

where \( r_a = \frac{1}{1 + \frac{1}{{y_a}^2}} \)

and \( y_a = \frac{1}{{C_a}^2} - 1 \)/2

b. the interarrival time distribution corresponding to this value of \( \sigma \) is of the form:

\[
a(t) = (1 - r_a) u_o(t) + r_a \lambda' e^{-\lambda' t}
\]

where \( \lambda' = r_a \lambda \) and \( u_o(t) \) is the unit impulse function defined as

\[
u_o(t) = \begin{cases} \infty & \text{t=0} \\ \phi & \text{t≠0} \end{cases}
\]

Proof

The two roots of the quadratic equation (3.58) can be written as

\[
\sigma = \frac{-B}{2A} + \frac{\sqrt{B^2 - 4AC}}{2A}
\]

or

\[
\sigma = \frac{-B}{2A} - \frac{\sqrt{B^2 - 4AC}}{2A}
\]

As \( \kappa \to \infty \), \( \alpha \to \phi \) and the first of these expressions diverges.

The second root can be written as
Using the binomial theorem the term under the square root can be written as

\[(1 - 4AC/B^2)^{1/2} = 1 - \frac{2AC}{B^2} - \frac{2A^2C^2}{B^4} - \ldots\]

In view of this, \(\sigma\) can be expressed as:

\[
\sigma = -\frac{B}{2A} \left( \frac{2AC}{B^2} + \frac{2A^2C^2}{B^4} + \ldots \right)
\]

\[
= -\frac{C}{B} (1 + \frac{AC}{B^2} + \ldots)
\]

From (3.16) above

\[
-\frac{C}{B} = \frac{\rho (a_2^2 - a_1^2) + k \alpha_1 \rho (\rho a_2 + a_1)}{\rho (k a_1 + \alpha_2 - a_1) + (k-1)/k}
\]

which, on manipulation, becomes

\[
-\frac{C}{B} = \frac{k-2k\alpha_1 + k^2 \alpha_1^2 + k^2 \alpha_1 \rho - k^2 \alpha_1^2 \rho}{k^2 \alpha_1 + k-2k\alpha_1 + k/\rho - 1/\rho}
\]

where \(\alpha_1\) is given in (5) as
\[ a_1 = \frac{c_a^2 - 1}{2(c_a^2 + 1)} + \frac{2}{k(c_a^2 + 1)} + \frac{\sqrt{(c_a^2 - 1)^2 + 8(c_a^2 - 1)/k + 8(1 - c_a^2)/k^2}}{2(c_a^2 + 1)} \]

Differentiating w.r.t \( k \), we get

\[ \frac{3}{\partial k} (-C/B) = ((k^2 a_1 + k - 2k a_1 + k/\rho - 1/\rho)(1 - 2a_1 + 2ka_1^2 + \n\frac{2k a_1 \rho (1 - a_1)}{\partial k} (k^2 (\rho + 2a_1 (1 - \rho)) - 2k)) \]

\[ + (k - 2ka_1 + k^2 a_1^2 + k^2 a_1 \rho - k a_1 \rho)(1 - 2a_1 + 2ka_1^2 + 1/\rho + \n+ \frac{3}{\partial k} (k^2 - 2k))/B^2 \]

It can be easily verified that this expression is positive, indicating that \(-C/B\) is monotone increasing as a function of \( k \), by the mean value theorem of differential calculus.

Writing \( x = -C/B \), \( \sigma \) can be expressed as

\[ \sigma = x(1 + (-\frac{A}{B})x + \ldots \) \]

Differentiating w.r.t \( x \), we get

\[ \frac{3}{\partial x} \sigma = 1 + 2(-\frac{A}{B})x + x^2 \frac{3}{\partial x} (-\frac{A}{B}) \]

which is again positive, indicating that \( \sigma \) is monotone increasing as a function of \( x \) and therefore as a function of \( k \).

Taking the limit of \( \sigma \) as \( k \to \infty \), we get
\[ \lim_{k \to \infty} \sigma = -\frac{C}{B} \]

and

\[ \lim_{k \to \infty} \left( -\frac{C}{B} \right) = 1-a_2(1-\rho) \]

and

\[ \lim_{k \to \infty} a_1 = \frac{(C^2 - 1)}{C^2 + 1} \quad \text{giving} \quad \lim_{k \to \infty} a_2 = \frac{2}{C^2 + 1} = \frac{1}{1+y_a} = r_a \]

Therefore \[ \lim_{k \to \infty} \sigma = 1-r_a(1-\rho) \]

which is the largest value (upper bound) for \( \sigma \) in the interval \((0,1)\).

Using (3.56)

\[ \lim_{k \to \infty} A^*(\theta) = \lim_{k \to \infty} \left[ a_1 + \frac{a_2^2 \lambda}{a_2 \lambda + \theta} \right] \]

\[ = (1-r_a) + \frac{r_a^2 \lambda}{r_a \lambda + \theta} \]

Defining \( \lambda' = r_a \lambda \), and inverting we get

\[ a(t) = (1-r_a)u_0(t) + r_a \lambda' e^{-\lambda't} \]

which completes the proof.

Q.E.D.
Box 3.1

(i) For the M/G/1 model, the maximum entropy solution is given by

\[ P_n = \begin{cases} (1-\rho) & n=0 \\ (1-\rho)g^nx & n>0 \end{cases} \]

where \( g = 1/(1+y_s) \), \( x = \frac{\rho(1+y_s)}{1+\rho y_s} \).

(ii) The maximum entropy solution for the M/G/1 model corresponds to the steady state solution of the underlying Markov process when the service time distribution is of the generalised exponential (GE) type:

\[ f(t) = (1-r_s)u_o(t) + r_s\mu' \exp\{-\mu' t\} \]

where \( r_s = 1/(1+y_s) \), \( \mu' = r_s\mu \) and \( u_o(t) \) is the unit impulse function.
For the G/M/1 model the maximum entropy solution is

\[
   p_n = \begin{cases} 
   (1-\rho) & n=0 \\
   (1-\rho)g x^n & n>0 
   \end{cases}
\]

where \( g = \frac{\rho}{\rho+y_a} \) , \( x = \frac{\rho+y_a}{1+y_s} \)

This solution corresponds to the generalised exponential (GE) type distribution.
3.6 Summary

Maximum entropy formalism has been used to analyse the $M/G/1$ and $G/M/1$ - queueing systems in equilibrium. The results are summarised in Boxes 3.1 and 3.2.

The maximum entropy solution for the $M/G/1$ has been derived subject to the minimum possible assumptions, without committance to a particular type of service time distribution and without directly involving the associated imbedded Markov process. However, it has been shown that the maximum entropy solution corresponds to the steady state solution of the associated imbedded Markov process [KLEI 75] when the service time distribution is of the generalised exponential (GE) type. It has also been shown that the equilibrium solution for any two-stage $M/G/1$ converges to the maximum entropy solution as the service time distribution converges to the GE-type distribution.

For the $C/M/1$, the maximum entropy solution is identical in form to the solution obtained using imbedded Markov process theory. However, it has been shown that a GE-type interarrival time distribution produces a solution that approximately maximises the entropy over all solutions of the two-stage $G/M/1$.

In the following chapter, these results will be extended and exploited to give useful results for general single server queues.
CHAPTER IV

SPECTRAL METHODS AND MAXIMUM ENTROPY IN THE ANALYSIS
OF THE G/G/1 QUEUEING SYSTEM

4.1 Introduction

The G/G/1 model represents an infinite capacity queueing system with arbitrary, independent and identically distributed (i.i.d.) service and inter-arrival times. Up to the present, the analysis of this system has proved to be quite obscure and only a few results exist in abstract form [KLEI 75a]. The probability distribution for the number of customers in the system is not known and all the performance metrics are obtainable only through simulation or approximation techniques. Even if these measures can be obtained, they are expected to depend on the G-type model chosen for the service or inter-arrival times.

In this chapter, the spectral solution of Lindley's integral equation [LIND 52] and maximum entropy are used in an attempt to overcome these problems. Lindley's integral equation is a Wiener-Hopf type equation defining the waiting time distribution as the steady state solution of an embedded Markov process. The spectral solution of this equation is due to Smith [SMIT 53].

In Section (4.2) some useful results are obtained for different variations of the G/G/1 using the spectral solution of Lindley's equation, with the help of Rouche's theorem.

In Section (4.3) a maximum entropy model for general queues is introduced and used to produce probability distributions for the number of customers in some types of the G/G/1. The implication of the maximum entropy results on operational analysis are discussed in section (4.4). Finally, the flow in the G/GE/1 and the GE/GE/1 is studied in section (4.5).
4.2 Exact Results for Some Variations of the G/G/1

In this section an embedded Markov process defined on the waiting time is used to produce exact results for important variations of the G/G/1. The main tools employed are Rouche's theorem and the spectrum factorisation of Lindley's integral equation. It is illustrated how the GE-type distribution derived in the previous chapter simplifies the analysis and leads to a set of satisfactory and interesting results.

Fundamental results from queueing theory are given in Section (4.2.1) followed in Section (4.2.2) by analysis of the G/GE/1 queueing system. The GE/G/1 and the G/G/1 with rational Laplace transform are considered in Sections (4.2.3) and (4.2.4) respectively. Numerical comparisons are given in Section (4.2.5).
4.2.1 Fundamental queuing theory results

Consider a stationary G/G/1 queueing system. Let \( t_{n+1} \) be the time between the \( n \)th and the \((n+1)\)th arrivals in this queue, where \( t_{n+1} \) is drawn from the PDF \( A(t) \) and \( E(t_{n+1}) = 1/\lambda \). Let \( S_n \) be the service time of the \( n \)th customer, where \( S_n \) is drawn from the PDF \( F(t) \) and \( E(S_n) = 1/\mu \).

Define \( u_n = S_n - t_{n+1} \) and let \( C(t) \) be the PDF of \( u_n \). Then, if \( W_n \) is the waiting time (in queue) of the \( n \)th customer with PDF \( W(t) \), we get

\[
W_{n+1} = \max(0, W_n + u_n) \tag{4.1}
\]

Lindley [LIND 52] has shown that the PDF for the random variable \( W_{n+1} \) is given by

\[
W(t) = \int_{-\infty}^{t} W(t-u) \, d(C(u)) \quad t > 0 \tag{4.2}
\]

which is valid for \( \lambda/\mu < 1 \).

A direct solution of (4.2) may not be feasible, but using spectrum factorisation [KLEI 75], the following results can be obtained from (4.2):

The Laplace transform of \( W(t) \) is given by:

\[
W*(\theta) = \theta \phi_+(\theta) \tag{4.3}
\]

where

\[
\phi_+(\theta) = K/\psi_+(\theta) \tag{4.4}
\]

and \( \psi_+(\theta) \) is obtainable from the factorisation

\[
A*(-\theta) F*(\theta) - 1 = \psi_+(\theta)/\psi_-(\theta) \tag{4.5}
\]

\( A*(\theta), F*(\theta) \) being the Laplace transforms of the interarrival and service time distributions respectively.
The factorisation is carried out such that $\psi_+(\theta), \psi_-(\theta)$ satisfy the following conditions:

(i) $\psi_+(\theta)$ is an analytic function of $\theta$ with no zeros in the half-plane $\text{Re}(\theta) > 0$.

(ii) $\psi_-(\theta)$ is an analytic function of $\theta$ with no zeros in the half-plane $\text{Re}(\theta) < D$ where $D$ is a constant ($D>0$).

Furthermore, if $A^*(\theta), F^*(\theta)$ are not to include any discontinuities and to be of finite moments, then Kleinrock [KLEI 75a] added the following two properties:

(iii) For $\text{Re}(\theta) > 0$, $\lim_{|\theta| \to \infty} \frac{\psi_+(\theta)}{\theta} = 1$ 

(iv) For $\text{Re}(\theta) < D$, $\lim_{|\theta| \to \infty} \frac{\psi_-(\theta)}{\theta} = -1$ 

(4.6)

However, if $A^*(-\theta), F^*(-\theta)$ are allowed to take jumps at the origin (unit impulse function), then conditions (4.6) may be relaxed. In such a case the only required condition will be

$$\lim_{|\theta| \to \infty} \frac{\psi_+(\theta)}{\psi_-(\theta)} = \text{const.}$$

The constant $K$ is given by

$$K = \lim_{\theta \to 0} \frac{\psi_+(\theta)}{\theta}$$

(4.7)

It has also been shown [MARS 68] that $K$ is actually the equilibrium probability that an arbitrary arrival finds an empty system. Denoting this probability as $\Pi_0$ we can write

$$\Pi_0 = K$$

(4.8)
Marshall [MARS 68] used the above results to generalise the Pollaczek-Khinchin (P-K) transform to give:

\[ W = \Pi_0 (1 - I(-\theta))/(1 - A(-\theta)F(\theta)) \]

where \( I(\theta) \) is the Laplace transform of the idle time distribution.

Furthermore, it has been shown by Kleinrock [KLEI 75a], using Kingman algebra for queues that the average waiting time in the G/G/1 is given by

\[ W = \frac{a_n^2 + a_f^2 + (\bar{a})^2(1-\rho)^2}{2\bar{a}(1-\rho)} - \frac{\overline{I^2}}{2I} \]

where \( a_n^2 \) and \( a_f^2 \) are the variances of the inter-arrival and service time distributions respectively. \( \bar{a} \) is the first moment of the inter-arrival time distribution, \( \overline{I} \) is the first moment of the idle time distribution, and \( \overline{I^2} \) is the second moment of the idle time distribution.

4.2.2 Analysis of the G/GE/1 queueing system

Although the results of section (4.2.1) have been available in the literature for some time now, analysts have so far refrained from using them in practice. This is mainly because of the difficulties arising in obtaining the factorisation (4.5). In this section it is illustrated how these difficulties can be overcome by use of Rouche's theorem and the generalised exponential distribution (GE-type) derived in Chapter 3. In later sections, the procedure is generalised to other variations of the G/G/1.

The system to be considered is a G/GE/1 with arbitrary (i.i.d.) inter-arrival times and a service time distribution of the generalised exponential type (GE):
\[ f(t) = (1-r_s)u_0(t) + r_s \mu' \exp(-\mu't) \]

where \( \mu' = r_s \mu \) and \( u_0(t) \) is the unit impulse function (see Chapter 3, theorem 3.2).

The Laplace transform of the above function is:

\[ F*(\theta) = (1-r_s) + r_s \mu'/(\mu'+\theta) \]  \hspace{1cm} (4.11)

Taking \( A*(\theta) \) to be the Laplace transform of the inter-arrival time distribution and substituting (4.11) in (4.5) we get:

\[ \psi_+/(\psi_-) = (((1-r_s)\theta+\mu')A*(\theta)-\mu'-\theta)/(\mu'+\theta) \]  \hspace{1cm} (4.12)

In order to factorise (4.12) the roots of the numerator in this equation should be determined. This can be achieved by studying the zeroes of the function

\[ \mu'+\theta-((1-r_s)\theta+\mu')A*(-\theta) = 0 \]  \hspace{1cm} (4.13)

The poles due to \( A*(-\theta) \) can be ignored since they all lie in the region \( \text{Re}(\theta)>0 \) and hence are of no interest in determining \( \psi_+/(\psi_-) \) \cite{KLEI 75a}. This is due to the fact that \( A(t) = 0 \) for \( t<0 \).

Clearly, one root of equation (4.13) occurs at \( \theta=0 \). The remaining roots are established by the following theorem:

**Theorem (4.1)**

For \( \rho \) in the open interval \((0,1)\) and \( r_s \) in the interval \([0,2)\), the polynomial (4.13) has only one zero inside the closed contour \( C \) defined on the negative half-plane \( \text{Re}(\theta) \leq 0 \) and shown in Fig. 4.1.

**Proof**

This result has been proved by Kleinrock \cite{KLEI 75a} for the case \( r_s=1 \). Here the proof is extended to include all the value of \( r_s \) in \([0,2)\).
Define
\[ h(\theta) = \mu^t + \theta \]
\[ g(\theta) = \{(1-r_s)\theta + \mu^t\} A^*(-\theta) \]

and choose C to be a contour forming a semicircle with an infinite radius given in Fig. 4.1. This contour is considered in an attempt to identify all the poles and zeros in the region \( \text{Re}(\theta) \leq 0 \) which can be properly included in the determination of \( \psi(\theta) \). Information on the number of zeros in \( \text{Re}(\theta) \leq 0 \) is obtained by making use of Rouche's theorem: "If \( h(\theta) \) and \( g(\theta) \) are analytic functions of \( \theta \) inside and on a closed contour \( C \), and if \( |g(\theta)| < |h(\theta)| \) on \( C \), then \( h(\theta) \) and \( h(\theta) + g(\theta) \) have the same number of zeros inside \( C \)."

By definition
\[ A^*(-\theta) = \int_0^\infty \exp(\theta t) \, dA(t) \]

and for \( \text{Re}(\theta) < 0 \) it is implied that \( \exp(\text{Re}(\theta) t) < 1 \) \( (t>0) \) and so, for \( i = \sqrt{1} \),

\[ |g(\theta)| = |\{(1-r_s)\theta + \mu_i\} \int_0^\infty \exp(\theta t) \, dA(t)| \]
\[ = |\{(1-r_s)\theta + \mu_i\} \int_0^\infty \exp(\text{Re}(\theta) t) \exp(i\text{Im}(\theta) t) \, dA(t)| \]
\[ < |\{(1-r_s)\theta + \mu_i\} \int_0^\infty \exp(i\text{Im}(\theta) t) \, dA(t)| \]

but
\[ |\exp(i\text{Im}(\theta))| = 1 \quad \text{and so} \]
\[ |g(\theta)| < |\{(1-r_s)\theta + \mu_i\} \int_0^\infty \, dA(t)| \]
\[ = |(1-r_s)\theta + \mu_i| \]
Fig. 4.1 The contour $C$ for the $G/GE/1$ queueing system.

Fig. 4.2 The excursion around the origin.
Therefore to prove that $|h(\theta)| > g(\theta)$ one need only prove that

$$|\mu'+\theta| > |(1-r_s)\theta+\mu'|$$  \hspace{1cm} (4.14)$$

Defining $\sigma = \text{Re}(\theta)$ and $\omega = \text{Im}(\theta)$, $\theta$ can be expressed as $\theta = \sigma + i\omega$ and consequently,

$$|\mu'+\theta| = \sqrt{(\sigma+\mu')^2 + \omega^2}$$

$$|\mu' + (1-r_s)\theta| = \sqrt{(\mu' + (1-r_s)\sigma)^2 + (1-r_s)^2 \omega^2}$$

If $|\mu'+\theta| > |\mu' + (1-r_s)\theta|$, then

$$((\sigma+\mu')^2 + \omega^2) > ((\mu' + (1-r_s)\sigma)^2 + (1-r_s)^2 \omega^2)$$

This will be the case if

$$\sigma^2 + \omega^2 = r_s(2-r_s)\omega^2 > (\mu' + (1-r_s)\sigma)^2 - (\mu'+\sigma)^2$$

$$= -2r_s\sigma\mu' - \sigma r_s(2-r_s)$$  \hspace{1cm} (4.15)$$

which is true if

$$-\frac{\sigma}{\sigma^2 + \omega^2} < \frac{(2-r_s)}{2r_s\mu}$$

Since the closed contour C is chosen such that $|\theta|$ is infinite, the LHS of the above inequality is equal to zero, which implies that the inequality is satisfied for $0 < r_s < 2$. For $r_s = 2$, the RHS becomes zero as well. However, it can be shown from (4.15) that the inequality (4.14) is not satisfied but the inequality $|(1-r_s)\theta + \mu'| > |\mu' + \theta|$ holds instead. Therefore, in this
case, one has to consider the particular polynomial.

This means that the conditions of Rouche's theorem are satisfied for \( \theta = 0 \) and \( r_s < 2 \). Note that \( |g(0)| = |h(0)| = \mu' \) at \( \theta = 0 \), which violates the conditions for Rouche's theorem. To overcome this problem, the contour \( C \) is allowed to make a small semicircular excursion of radius \( \varepsilon \) (\( \varepsilon > 0 \)) to the left of the origin, as shown in Fig. 4.2.

Consider an arbitrary point \( \theta \) on this semicircle which lies at an angle \( \alpha \) with the Re(\( \theta \)) axis. Expressing this point as

\[
\theta = \text{Re}(\theta) + i \text{Im}(\theta)
\]

\[
= -\varepsilon \cos \alpha + i \varepsilon \sin \alpha
\]

it is implied that

\[
|h(\theta)|^2 = |\theta + \mu'|^2
\]

\[
= |-\varepsilon \cos \alpha + i \varepsilon \sin \alpha + \mu'|^2
\]

\[
= \mu'^2 - 2\mu' \varepsilon \cos \alpha + O(\varepsilon^2)
\]

Evaluating \( g(\theta) \) on this same semicircular excursion, it follows that

\[
|g(\theta)|^2 = |(1-r_s)\theta + \mu'|^2 \int_0^\infty \exp\{(-\varepsilon \cos \alpha + i \varepsilon \sin \alpha)t\} dA(t)|^2
\]
\[ (1-r_s) \theta + \mu' \mid^2 (1-\varepsilon \cos \alpha / \lambda + i \varepsilon \sin \alpha / \lambda + O(\varepsilon^2)) \]
\[ = \mu'^2 - 2\mu'(1-r_s)\varepsilon \cos \alpha + O(\varepsilon^2) \mid 1-2\varepsilon \cos \alpha / \lambda + O(\varepsilon^2) \]
\[ = \mu'^2 - 2\mu'(1-r_s)\varepsilon \cos \alpha / \lambda - 2\mu'(1-r_s)\varepsilon \cos \alpha + O(\varepsilon^2) \]

Therefore, we need only prove the inequality:
\[ \{ \mu'^2 - 2\mu' \varepsilon \cos \alpha \} > \{ \mu'^2 - 2\mu'^2 \varepsilon \cos \alpha / \lambda - 2\mu'(1-r_s)\varepsilon \cos \alpha \} \]  
(4.16)

which becomes on manipulation:
\[ \{ 2\mu'^2 \varepsilon \cos \alpha / \lambda > 2r_s \mu' \varepsilon \cos \alpha \} \] or \( \mu / \lambda > 1 \)

which means that the inequality (4.16) is satisfied for \( \rho < 1 \).

Since \( h(\theta) \) has only one zero at \( \theta = -\mu' \), it follows by Rouche's theorem, that \( h(\theta) + g(\theta) \) has only one zero inside the contour \( \text{Re}(\theta) \leq 0 \), for \( \rho < 1 \).

Q.E.D.

Let \( \theta = 0 \) and \( \theta = -\theta_1 \), the two zeros in the region \( \text{Re}(\theta) \leq 0 \).

Expression (4.12) can be now rewritten as
\[ \frac{\psi_+(\theta)}{\psi_-(\theta)} = \left\{ \frac{(1-r_s)\theta + \mu'}{\mu' + \theta} \right\} \left\{ \frac{A^*(-\theta) - \mu' - \theta}{\theta + \theta_1} \right\} \]  
(4.17)

where the first bracketed term contains no poles and zeros in \( \text{Re}(\theta) < 0 \).

The term \( \text{Re}(\theta) < 0 \) is extended to the region \( \text{Re}(\theta) < D \), where \( D > 0 \) is chosen such that no new zeros or poles of equation (4.13) are introduced.
The first bracket qualifies for $\psi_-(\theta)^{-1}$. It follows immediately that the second bracket qualifies for $\psi_+(\theta)$ since some of its zeros ($\theta=0, \theta=\theta_1$) or poles ($\theta=\mu'$) are in the region $\text{Re}(\theta) > 0$. Thus, equation (4.17) is factorised into the following form:

$$
\psi_+(\theta) = \frac{\theta(\theta+\theta_1)}{\mu'+\theta} \quad (4.18)
$$

$$
\psi_-(\theta) = \frac{-\theta(\theta+\theta_1)}{(\mu'+\theta-(1-r_s)(\theta+\mu')A*(-\theta))} \quad (4.19)
$$
4.2.2a The waiting time distribution and the average number of customers in the system

Theorem 4.2

For the G/E/1 queueing system described above, the waiting time distribution is given by:

\[ w(t) = \left( \frac{\theta_1}{\mu} \right) u_0(t) + (1 - \theta/\mu') \theta_1 e^{-\theta_1 t} \]  
\[ (4.20) \]

The mean waiting time is

\[ W = \frac{1}{\theta_1} - \frac{1}{\mu'} \]  
\[ (4.21) \]

The mean response time is

\[ T = \frac{1}{\theta_1} - \frac{y_s}{\mu} \]  
\[ (4.22) \]

The average number of customers in the system is

\[ \langle n \rangle = \frac{\lambda}{\theta_1} - \rho y_s \]  
\[ (4.23) \]

Proof

Using (4.3) and (4.4)

\[ W^*(\theta) = K \theta/\psi_+(\theta) \]

where \( K = \lim_{\theta \to 0} \psi_+(\theta)/\theta \)

But by (4.18) \( \psi_+(\theta) = \frac{\theta(\theta + \theta_1)}{\mu' + \theta} \)

which gives

\[ K = \frac{\theta_1}{\mu'} \]

and consequently
\[ W^*(\theta) = \frac{\theta_1 (\mu' + \theta)}{\mu' (\theta_1 + \theta)} = \frac{(\theta_1 / \mu') (\mu' + \theta)}{\theta_1 + \theta} = \frac{\theta_1 + (\theta_1 / \mu') \theta}{\theta_1 + \theta} \]

which, on inversion, yields (4.20).

The mean waiting time is obtained by performing the integral

\[ W = \int_0^\infty t w(t) \, dt = \frac{1}{\theta_1} - \frac{1}{\mu} \]

which is (4.21).

From queueing theory it is known:

\[ T = W + \frac{1}{\mu} = \frac{1}{\theta_1} - \frac{y_s}{\mu} \]

Applying Little's formula

\[ <n> = \lambda T \]

\[ = \frac{\lambda}{\theta_1} - \rho y_s \]

Q.E.D.

Examples

1. The H2/GE/1

For this system:

\[ A^*(\theta) = \frac{\alpha_1 \lambda_1}{\lambda_1 + \theta} + \frac{\alpha_2 \lambda_2}{\lambda_2 + \theta} \]

where \( \lambda_1 = 2\alpha_1 \lambda \) and \( \lambda_2 = 2\alpha_2 \lambda \)

where \( \alpha_1 \) is given by (3.41b)

in view of which polynomial (4.13) becomes

\[ \theta^2 + \theta \{ \mu' - \lambda_1 \lambda_2 / \lambda - r_s \alpha_1 \lambda_1 + \alpha_2 \lambda_2 \} - \frac{\mu \lambda_1 \lambda_2}{\lambda} (1-\rho) = 0 \]
Defining

\[ B = \frac{\lambda_1 \lambda_2}{\lambda} + r_s (a_1 \lambda_1 + a_2 \lambda_2) - \mu' \]

the roots of the above polynomial are

\[ \theta = \frac{B}{2} + \frac{\sqrt{B^2 + 4\mu' \lambda_1 \lambda_2 (1-\rho) / \lambda}}{2} \]

It can be easily verified that these roots are of opposite sign for \( \rho < 1 \), and the required zero is

\[ \theta_1 = -\theta = -\frac{B}{2} + \frac{\sqrt{B^2 + 4\mu' \lambda_1 \lambda_2 (1-\rho) / \lambda}}{2} \]

where we have chosen the root \( \theta = \frac{B}{2} - \frac{\sqrt{B^2 + 4\mu' \lambda_1 \lambda_2 (1-\rho) / \lambda}}{2} < 0 \)

2) The \( E_2 / G E / I \)

For this system, \( A^*(\theta) = \frac{4\lambda^2}{(2\lambda + \theta)^2} \)

on the basis of this the polynomial (4.13) becomes

\[ \theta^2 - \theta (4\lambda - \mu') - 4\lambda \mu' (1-\rho) = 0 \]

the roots of which are

\[ \theta = \frac{4\lambda - \mu'}{2} + \frac{\sqrt{(4\lambda + \mu')^2 - 16\lambda^2 r_s}}{2} \]

It can be verified that these roots are of opposite sign for \( \rho < 1 \) and the required zero is

\[ \theta_1 = -\theta = -\frac{(4\lambda - \mu')}{2} + \frac{\sqrt{(4\lambda + \mu')^2 - 16\lambda^2 r_s}}{2} \]

where we have chosen the negative sign in the above expression for \( \theta \).

For numerical results, see Tables 4.1-4.
4.2.3 The average number of customers in the GE/G/1

Having analysed the G/GE/1, we now turn to its dual - the GE/G/1 - which is a queueing system with a GE-type interarrival time distribution and a G-type (arbitrary) service time distribution. In this section it is shown that under wide circumstances, the average number of customers in the GE/G/1 system is unaffected by the type of the service time distribution.

To obtain the average number of customers in the GE/G/1 use is made of the mean value formula for the waiting time in the G/G/1 queueing system, given by (4.10) as:

\[ W = \frac{\sigma_a^2 + \sigma_f^2 + (\bar{a})(1-\rho)^2}{2a(1-\rho)} - \frac{I^2}{2I} \]  \hspace{1cm} (4.24)

where \( \sigma_a^2, \sigma_f^2 \) are the variances of the interarrival and service time distributions, respectively. \( \bar{a} \) is the first moment of the interarrival time distribution, \( I \) is the first moment of the idle time distribution, and \( I^2 \) is the second moment of the idle time distribution.

Using the substitutions \( \bar{a} = 1/\lambda, \sigma_a^2 = C_a^2/\lambda, \sigma_f^2 = C_s^2/\mu^2 \) and knowing that \( <n> = \lambda W + \rho \) (Little's formula), we get:

\[ <n> = \frac{1 + y_a + \rho^2 y_s}{(1-\rho)} - \frac{\lambda I^2}{2I} \]  \hspace{1cm} (4.25)

The first part of this formula is standard for any G/G/1. The second part consists of the ratio \( \lambda I^2/2I \) which varies with the type of system.

Therefore, to investigate the effect of the service time distribution on the mean value formula (4.25), one need only consider the effect on the ratio \( \lambda I^2/2I \). This ratio can be expressed in terms of \( \psi_+(0), \)
using Marshall's generalisation of the P-K transform, which is given by (4.9). From (4.9) it can easily be deduced that

$$ I^*(-\theta) = 1 + \theta/\psi_-(\theta) $$

(4.26)

Since $$ \bar{I} = \frac{dI^*(-\theta)/d\theta}{\theta=0} $$ and $$ \bar{I}^2 = \frac{d^2 I^*(-\theta)/d\theta^2}{\theta=0} $$, one immediately gets

$$ \frac{\bar{I}^2}{\bar{I}} = -\frac{\psi'_-(\theta)}{\psi_-(\theta)} $$

(4.27)

where $$ \psi'_-(\theta) = \frac{d\psi_-(\theta)}{d\theta} $$.

To determine $$ \psi_-(\theta) $$, consider the case where the G-type service time distribution has a rational Laplace transform i.e. $$ F^*(\theta) $$ can be expressed as:

$$ F^*(\theta) = \frac{NF^*(\theta)}{DF^*(\theta)} $$

(4.28)

In this case (4.5) gives:

$$ \frac{\psi_+(\theta)}{\psi_-(\theta)} = \frac{NF^*(\theta)A^*(-\theta) - DF^*(\theta)}{DF^*(\theta)} $$

(4.29)

To factorise the RHS of this expression one need only consider the zeros of the expression

$$ DF^*(\theta) - NF^*(\theta)A^*(-\theta) = 0 $$

(4.30)

In view of this discussion the following theorem can be presented.
Theorem (4.3)

For any GE/G/1, if:

i) The service time distribution has a rational Laplace transform.

ii) The degree of polynomial $NF^*(\theta)$ is less than or equal to the degree of $DF^*(\theta)$.

iii) The polynomials $DF^*(\theta)$ and $\{NF^*(\theta)A^*(-\theta)\}$ are both analytic functions of $\theta$ inside and on a closed contour $C$ defined on the region $\text{Re}(\theta) < 0$.

iv) $|DF^*(\theta)| > |NF^*(\theta)A^*(-\theta)|$ on $C$

then

$$\overline{I^2}/2I = 1/\lambda' \quad \text{where} \quad \lambda' = r\lambda \quad (4.31)$$

and

$$<n> = \frac{\rho(1+y_a+py_s)}{(1-\rho)} \quad \left( \text{if} \quad \theta > \frac{1-C_s^z}{1+C_s^z} \right) \quad (4.32)$$

Proof

Let $C$ be the closed contour with infinite radius defined in the previous section and shown in Fig.4.1. If conditions (iii) and (iv) are satisfied, then by Rouche's theorem, the polynomial (4.30) and $DF^*(\theta)$ has the same number of zeros inside the contour $C$. Condition (ii) will ensure that the zeros of the polynomial (4.30) can exceed those of $DF^*(\theta)$ by at most one zero. Since $\theta = 0$ is always a zero, and $DF^*(\theta)$ has no zeros in $\text{Re}(\theta) > 0$ ($F(t) = 0$ for $t < 0$), then the polynomial (4.30) has no zeros in $\text{Re}(\theta) > 0$.

Consequently, (4.29) can be factorised in the following manner:
\[
\frac{\psi_+(\theta)}{\psi_- (\theta)} = \frac{DF*(\theta)(\lambda'-\theta)-NF*(\theta)(\lambda'(1-r_a)\theta)}{DF*(\theta)} \{ \text{CONST} \} \frac{1}{\lambda' - \theta}
\]

which implies that \( \psi_-(\theta) = \frac{(\lambda'-\theta)}{\text{CONST}} \)

\[
\psi_-'(\theta) = -1/\text{CONST}
\]

\[
\frac{\lambda'}{2I} = \frac{1}{r_a} = \frac{C_a^2 + 1}{2} = 1 + y_a
\]

Taking this into (4.25) one gets

\[
\langle n \rangle = \rho \left( 1 + y_a + py_s \right) / (1 - \rho)
\]

Q.E.D.

The conditions of theorem (4.3) are believed to apply to a wide variation of the G-type service type distribution, which indicates that the mean value formula (4.32) is largely invariant to the type of distribution function used for the service time. All that matters is the first and second moment of the service time.

To illustrate, consider the following examples.

1) The \( \text{GE/H}_2/1 \)

For this system

\[
F*(\theta) = \frac{\alpha_1 \mu_1}{\mu_1 + \theta} + \frac{\alpha_2 \mu_2}{\mu_2 + \theta} = \frac{\mu_1 \mu_2 + \alpha_1 \mu_1 \theta + \alpha_2 \mu_2 \theta}{\mu_1 \mu_2 + \mu_1 \theta + \mu_2 \theta + \theta^2}
\]

\[
\therefore \quad NF*(\theta) = \mu_1 \mu_2 + (\alpha_1 \mu_1 + \alpha_2 \mu_2) \theta
\]

\[
DF*(\theta) = \mu_1 \mu_2 + (\mu_1 + \mu_2) \theta + \theta^2
\]
Consequently, the polynomial (4.30) becomes

\[(\mu_1 \mu_2 + (\mu_1 + \mu_2) \theta + \theta^2) - (\mu_1 \mu_2 + (a_1 \mu_1 + a_2 \mu_2) \theta) A^*(-\theta) = 0\]

It can be easily verified that Rouche's theorem applies to this polynomial (conditions (iii) and (iv) of Th. (4.3)), implying that

\[\psi_-(\theta) = (\lambda' - \theta)/\text{const}\]

and consequently

\[<n> = \frac{\rho(1+y_a + py_s)}{1-\rho}\]

2) \ The \ GE/E_k/1

For this system

\[F^*(\theta) = \left(\frac{k\mu}{k\mu+\theta}\right)^k\]

\[NF^*(\theta) = (k\mu)^k\]

\[DF^*(\theta) = (k\mu+\theta)^k\]

The polynomial (4.30) becomes

\[(k\mu+\theta)^k - (k\mu)^k A^*(-\theta) = 0\]

It can again be verified that this expression satisfies Rouche's theorem, implying that \[\psi_-(\theta) = (\lambda' - \theta)/\text{const}\]

\[<n> = \frac{\rho(1+y_a + py_s)}{1-\rho}\]

\[\therefore\]
4.2.4 Other Variations of the G/G/1 System

In this section, the general case where both service and interarrival times are arbitrary but with rational Laplace transforms is considered. The main motivation is to obtain exact results for some important variations of the G/G/1 and investigate the effect of the service time on \( \langle n \rangle \) when the interarrival time is not of the CE type. In this case both \( A^*(\theta) \) and \( F^*(\theta) \) can be expressed as:

\[
A^*(\theta) = \frac{NA^*(\theta)}{DA^*(\theta)}
\]

\[
F^*(\theta) = \frac{NF^*(\theta)}{DF^*(\theta)}
\]

For this system the following theorem can be presented:

**Theorem (4.4)**

For any G/G/1 system, if:

i) Both service and interarrival times have rational Laplace transforms.

ii) The polynomials \( DF^*(\theta) \) and \( \{NF^*(\theta)A^*(-\theta)\} \) are both analytic functions of \( \theta \) inside and on a closed contour \( C \) defined on \( \Re(\theta)<0 \).

iii) \( |DF^*(\theta)| > |NF^*(\theta)A^*(-\theta)| \) on \( C \)

then the average number of customers is given by

\[
\langle n \rangle = \lambda \left[ \frac{1 + y}{a + \rho y_s} \right] - \lambda \left\{ \sum_{k=i+2}^{m} \frac{DA^*(\theta)}{DA^*(\theta)} \theta_k + \sum_{j=i+2, j \neq k}^{m} \frac{m}{m} \theta_j \right\} \tag{4.34}
\]

where \( m \) is the total number of zeroes for the polynomial

\[
DF^*(\theta) - NF^*(\theta)A^*(-\theta) = 0 \tag{4.35}
\]

and \( i \) is the total number of zeroes for \( DF^*(\theta) \), and

\( \theta_{i+2}, \theta_{i+3}, \ldots, \theta_m \) are all in \( \Re(\theta)>0 \).
Proof

This is a generalisation of theorem (4.3). Let C be the closed contour defined in theorem (4.3). If (ii) and (iii) above are satisfied, then by Rouche's theorem \(DF^*(\theta)\) and \(\{DF^*(\theta)-NF^*(\theta)A^*(-\theta)\}\) will have the same number of zeroes inside C. Since \(\theta=0\) is always a zero, and \(DF^*(\theta)\) has no zeroes in \(Re(\theta)>0\) \(F(t)=0, t<0\), then the polynomial (4.35) will have only \(m-i-1\) zeroes in \(Re(\theta)>0\), where \(i\) and \(m\) are as defined above. Denoting these zeroes \(\theta_{i+2}, \theta_{i+3}, \ldots, \theta_m\), the factorisation (4.35) can be written as follows:

\[
\frac{\psi_+(\theta)}{\psi_-(\theta)} = \frac{DF^*(\theta)DA^*(-\theta) - NF^*(\theta)NA^*(-\theta)}{DF^*(\theta)(\theta-\theta_{i+2}) \ldots (\theta-\theta_m)} \left\{ \frac{(\theta-\theta_{i+2}) \ldots (\theta-\theta_m)}{DA^*(-\theta)} \right\}
\]

or, if all the roots are determined:

\[
\frac{\psi_+(\theta)}{\psi_-(\theta)} = \frac{\theta(\theta_{i+2}+\theta) \ldots (\theta_{i+3}+\theta)}{DF^*(\theta)} \left\{ \frac{(\theta-\theta_{i+2}) \ldots (\theta-\theta_m)}{DA^*(-\theta)} \right\} \quad (4.36)
\]

This gives \(\psi_-(\theta) = \frac{DA^*(-\theta)}{(\theta-\theta_{i+2}) \ldots (\theta-\theta_m)}\)

Differentiating, we get

\[
\psi_-'(\theta) = \frac{DA^*(\theta) \prod_{k=i+2}^{m} (\theta-\theta_k) + DA^*(-\theta) \sum_{k=i+2}^{m} \prod_{j=i+2, j \neq k}^{m} (\theta-\theta_j)}{\prod_{k=i+2}^{m} (\theta-\theta_k)^2}
\]

Using (4.27) and (4.29) one gets (4.35).

Q.E.D.

Therefore, if the roots \(\theta_{i+2}, \ldots, \theta_m\) can be located, the average number of customers in the system can easily be evaluated.
Examples

1) The $E_2/H_2/1$

For this system

$$A^*(\theta) = \left(\frac{2\lambda}{2\lambda+\theta}\right)^2$$

$$F^*(\theta) = \frac{\alpha_1\mu_1}{\mu_1+\theta} + \frac{\alpha_2\mu_2}{\mu_2+\theta}$$

\[\therefore \quad NA^*(-\theta) = 4\lambda^2, \quad DA^*(-\theta) = (2\lambda-\theta)^2\]

$$NF^*(\theta) = \mu_1\mu_2 + (\alpha_1\mu_1+\alpha_2\mu_2)\theta$$

$$DF^*(\theta) = \mu_1\mu_2 + (\mu_1+\mu_2)\theta^2$$

The polynomial to be considered is:

$$\theta^3 - \theta^2 \{4\lambda - \mu_1 - \mu_2\} + \theta \{4\lambda^2 - 4\lambda (\mu_1+\mu_2) + \mu_1\mu_2\} - 4\lambda\mu_1\mu_2(1-p) = 0$$

It can be verified that the conditions of theorem (4.4) apply implying that the above polynomial has two zeroes in $\text{Re}(\theta) < 0$ and only one zero, $\theta_4$, in $\text{Re}(\theta) > 0$. Therefore,

$$\psi_-(\theta) = -\frac{(2\lambda-\theta)^2}{\theta_4 - \theta}$$

$$\psi_-'(\theta) = \frac{4\lambda^2 - \theta^2 - 4\lambda\theta_4 + 2\theta\theta_4}{(\theta_4 - \theta)^2}$$

\[\therefore \quad \frac{\psi_-'(0)}{\psi_-(0)} = \frac{4\lambda(\lambda-\theta_4)}{\theta_4} \cdot \frac{\theta_4}{4\lambda^2} = \frac{\lambda-\theta_4}{\lambda\theta_4}\]
Consequently,
\[ \langle n \rangle = \frac{1+y_1+y_2}{1-p} + \frac{\lambda}{\theta} - 1 \]  

(4.37)

\( \theta_4 \) can easily be determined using the Newton-Raphson method. In Table 4.2 the mean value formula (4.37) is compared with that of the equivalent E\(_2\)/GE/1 and simulation.

2) The E\(_2\)/E\(_2\)/1

In this case
\[ A^*(\theta) = \left(\frac{2\lambda}{2\lambda+\theta}\right)^2, \quad F^*(\theta) = \left(\frac{2\mu}{2\mu+\theta}\right)^2 \]

The polynomial to be considered is
\[ \theta^3 + \theta^2 (\mu - \lambda) + 4\theta (\mu^2 - 4\lambda \mu + \lambda^2) - 16\lambda \mu (\mu - \lambda) = 0 \]

which has two zeroes in Re(\( \theta \))<0 and only one zero, \( \theta_4 \), in Re(\( \theta \))>0.

Equation (4.37) again applies with \( \theta_4 \) determined by the use of the Newton-Raphson method. In Table 4.3, the average number of customers for the E\(_2\)/E\(_2\)/1 is compared with that of the equivalent E\(_2\)/GE/1.

3) The H\(_2\)/H\(_2\)/1

In this case
\[ A^*(\theta) = \frac{\gamma_1 \lambda_1}{\lambda_1 + \theta} + \frac{\gamma_2 \lambda_2}{\lambda_2 + \theta}, \quad F^*(\theta) = \frac{\alpha_1 \nu_1}{\nu_2 + \theta} + \frac{\alpha_2 \nu_2}{\nu_2 + \theta} \]

The polynomial to be considered is
\[ \theta^3 + \theta^2 \left( \nu_1 + \nu_2 - \lambda_1 - \lambda_2 \right) + \theta \left[ \gamma_1 \lambda_2 \nu_1 \nu_2 - (\lambda_1 + \lambda_2) (\nu_1 + \nu_2) + (\alpha_1 \nu_1 + \alpha_2 \nu_2) (\gamma_1 \lambda_1 + \gamma_2 \lambda_2) \right] \\
- \frac{\nu_1 \nu_2 \lambda_1 \lambda_2}{\lambda} (1-p) = 0 \]
It can be verified that this polynomial has 2 zeroes in \( \text{Re}(\theta) < 0 \) and only one zero in \( \text{Re}(\theta) > 0 \).

Equation (4.35) becomes

\[
\langle n \rangle = \left( \frac{1 + y + \rho y}{1 - \rho} \right) + \frac{\lambda}{\theta_4} - \frac{\lambda(\lambda_1 + \lambda_2)}{\lambda_1 \lambda_2}
\]

where \( \theta_4 \) can easily be determined using the Newton-Raphson method. In Table 4.4 the \( H_2/H_2/1 \) is compared with the equivalent \( H_2/GE/1 \). Similar comparisons are carried out in Tables 4.5 and 4.6 for the case where the interarrival and service times are hypoexponential.

4.2.5 Numerical comparisons

Tables 4.1-4.5 show the average number of customers for different types of the \( G/G/1 \) system.

In Table 4.1, the average number of customers for the \( E_2/H_2/1 \) is compared with simulation, \( E_2/GE/1 \), \( GE/GE/1 \) and the diffusion approximation. The results appear to be reasonably close to each other. However, the results reveal that, even if we fix the inter-arrival time distribution, the service time has an effect on the average number of customers in the system. In Table 4.2 a similar comparison is carried out for low values of \( \rho \). The difference between the results increases in this case, but they are still reasonably close to one another.

These comparisons are continued for the \( E_2/E_2/1 \) and the \( H_2/H_2/1 \) in Tables 4.3 and 4.4 respectively. Again the results reveal the effect of the service and interarrival time pdf on the average number of customers in the system. The \( E_2/Hypo/1 \) is compared with the equivalent systems in Table 4.5. The \( \text{Hypo} \) here stands for an \( H_2 \) distribution with \(-1 < \alpha_1 < 0\) and \( \alpha_1 + \alpha_2 = 1 \) (see Chapter 3).
<table>
<thead>
<tr>
<th>$c_a^2$</th>
<th>$\rho$</th>
<th>$c_s^2$</th>
<th>$E_2/H_2/1$ ( \langle n \rangle ) SIM</th>
<th>$E_2/H_2/1$ ( \langle n \rangle ) exact</th>
<th>$E_2/GE/1$ ( \langle n \rangle ) exact</th>
<th>$GE/GE/1$ ( \langle n \rangle ) exact</th>
<th>$G/P$ ( \langle n \rangle ) D1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.75</td>
<td>2</td>
<td>3.44±0.05</td>
<td>3.440</td>
<td>3.421</td>
<td>3.375</td>
<td>3.56</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>5.67±0.12</td>
<td>5.684</td>
<td>5.652</td>
<td>5.625</td>
<td>5.813</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>10.08±0.32</td>
<td>10.18</td>
<td>10.139</td>
<td>10.125</td>
<td>10.313</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16</td>
<td>19.27±0.83</td>
<td>19.177</td>
<td>19.132</td>
<td>19.125</td>
<td>19.313</td>
</tr>
<tr>
<td></td>
<td></td>
<td>32</td>
<td>37.39±1.92</td>
<td>37.176</td>
<td>37.129</td>
<td>27.125</td>
<td>37.313</td>
</tr>
<tr>
<td></td>
<td></td>
<td>64</td>
<td>73.02±4.73</td>
<td>73.175</td>
<td>73.127</td>
<td>73.125</td>
<td>73.313</td>
</tr>
<tr>
<td></td>
<td></td>
<td>128</td>
<td>146.0±14</td>
<td>145.175</td>
<td>145.126</td>
<td>145.125</td>
<td>145.313</td>
</tr>
<tr>
<td>0.8</td>
<td></td>
<td>2</td>
<td>4.67±0.09</td>
<td>4.667</td>
<td>4.647</td>
<td>4.6</td>
<td>4.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>7.83±0.22</td>
<td>7.861</td>
<td>7.827</td>
<td>7.8</td>
<td>8.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16</td>
<td>27.24±1.39</td>
<td>27.054</td>
<td>27.008</td>
<td>27.00</td>
<td>27.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>32</td>
<td>52.95±3.02</td>
<td>52.653</td>
<td>52.604</td>
<td>52.6</td>
<td>52.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>64</td>
<td>102.4±8.0</td>
<td>103.853</td>
<td>103.802</td>
<td>103.800</td>
<td>104.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>128</td>
<td>203.7±21.0</td>
<td>206.252</td>
<td>206.201</td>
<td>206.2</td>
<td>206.4</td>
</tr>
</tbody>
</table>

Table 4.1 The average number of customers in the system compared for some variations of the G/G/1:

The $E_2/H_2/1$ versus simulation and other systems.
Table 4.2 The average number of customers compared for some type of the \( G/G/1 \):

The \( E_2/H_2/1 \) versus other systems

<table>
<thead>
<tr>
<th>( C_a^2 )</th>
<th>( \rho )</th>
<th>( C_s^2 )</th>
<th>( E_2/H_2/1 ) (&lt;n&gt;_{\text{exact}})</th>
<th>( E_2/GE/1 ) (&lt;n&gt;_{\text{exact}})</th>
<th>( GE/GE/1 ) (&lt;n&gt;_{\text{exact}})</th>
<th>( G/P ) (&lt;n&gt;_{\text{D}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>2</td>
<td>1.052</td>
<td>1.041</td>
<td>1.00</td>
<td>1.125</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>1.547</td>
<td>1.525</td>
<td>1.5</td>
<td>1.625</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>2.543</td>
<td>2.514</td>
<td>2.500</td>
<td>2.625</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16</td>
<td>4.541</td>
<td>4.507</td>
<td>4.500</td>
<td>4.625</td>
</tr>
<tr>
<td>0.25</td>
<td>2</td>
<td>0.326</td>
<td>0.322</td>
<td>0.291</td>
<td>0.354</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.402</td>
<td>0.395</td>
<td>0.375</td>
<td>0.438</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>0.569</td>
<td>0.554</td>
<td>0.541</td>
<td>0.604</td>
<td></td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>0.902</td>
<td>0.882</td>
<td>0.875</td>
<td>0.938</td>
<td></td>
</tr>
<tr>
<td>$c_a^2$</td>
<td>$\rho$</td>
<td>$c_s^2$</td>
<td>$E_2/E_2/1$ $&lt;n&gt;_{\text{exact}}$</td>
<td>$E_2/GE/1$ $&lt;n&gt;_{\text{exact}}$</td>
<td>$GE/GE/1$ $&lt;n&gt;_{\text{exact}}$</td>
<td>$G/P$ $&lt;n&gt;_{D}$</td>
</tr>
<tr>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>-------------------------------</td>
<td>-------------------------------</td>
<td>-------------------------------</td>
<td>----------------</td>
</tr>
<tr>
<td>0.25</td>
<td>0.25</td>
<td>0.273</td>
<td>0.274</td>
<td>0.229</td>
<td>0.292</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>0.695</td>
<td>0.701</td>
<td>0.625</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td>0.75</td>
<td>0.75</td>
<td>1.777</td>
<td>1.789</td>
<td>1.688</td>
<td>1.875</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.3 The average number of customers in the system compared for some types of the $G/G/1$:

The $E_2/E_2/1$ versus other systems
<table>
<thead>
<tr>
<th>$p$</th>
<th>$C_a^2$</th>
<th>$C_s^2$</th>
<th>$H_2/H_2/1\langle n\rangle_{\text{exact}}$</th>
<th>$H_2/GE/1\langle n\rangle_{\text{exact}}$</th>
<th>$GE/GE/1\langle n\rangle_{\text{exact}}$</th>
<th>$G/P\langle n\rangle_D$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>2</td>
<td>2</td>
<td>0.421</td>
<td>0.419</td>
<td>0.541</td>
<td>0.416</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>0.517</td>
<td>0.520</td>
<td>0.625</td>
<td>0.500</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>0.698</td>
<td>0.710</td>
<td>0.792</td>
<td>0.667</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16</td>
<td>1.043</td>
<td>1.068</td>
<td>1.125</td>
<td>1.00</td>
</tr>
<tr>
<td>0.75</td>
<td>2</td>
<td>2</td>
<td>0.475</td>
<td>0.471</td>
<td>0.875</td>
<td>0.500</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>0.592</td>
<td>0.598</td>
<td>0.958</td>
<td>0.583</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>0.800</td>
<td>0.827</td>
<td>1.125</td>
<td>0.750</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16</td>
<td>1.175</td>
<td>1.235</td>
<td>1.458</td>
<td>1.083</td>
</tr>
<tr>
<td>0.75</td>
<td>2</td>
<td>2</td>
<td>5.280</td>
<td>5.304</td>
<td>5.625</td>
<td>5.25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>7.579</td>
<td>7.647</td>
<td>7.875</td>
<td>7.500</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>12.117</td>
<td>12.228</td>
<td>12.375</td>
<td>12.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>7.469</td>
<td>7.532</td>
<td>8.625</td>
<td>7.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>9.858</td>
<td>10.039</td>
<td>10.875</td>
<td>9.75</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>14.777</td>
<td>14.35</td>
<td>15.375</td>
<td>14.25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16</td>
<td>23.857</td>
<td>23.537</td>
<td>4.375</td>
<td>23.25</td>
</tr>
</tbody>
</table>

Table 4.4 The average number of customers in the system compared for different types of the G/G/1:

The $H_2/H_2/1$ versus other systems.
<table>
<thead>
<tr>
<th>$c_a^2$</th>
<th>$\rho$</th>
<th>$c_s^2$</th>
<th>$&lt;n&gt;$ exact $E_2$/HYPO/1</th>
<th>$&lt;n&gt;$ exact $E_2$/GE/1</th>
<th>$&lt;n&gt;$ exact GE/GE/1</th>
<th>G/P</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.25</td>
<td>0</td>
<td>0.252</td>
<td>0.262</td>
<td>0.208</td>
<td>0.271</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.25</td>
<td>0.262</td>
<td>0.268</td>
<td>0.219</td>
<td>0.281</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>0.271</td>
<td>0.274</td>
<td>0.229</td>
<td>0.292</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.75</td>
<td>0.280</td>
<td>0.281</td>
<td>0.240</td>
<td>0.302</td>
</tr>
<tr>
<td>0.5</td>
<td>0.25</td>
<td>0</td>
<td>0.569</td>
<td>0.604</td>
<td>0.500</td>
<td>0.625</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.25</td>
<td>0.630</td>
<td>0.650</td>
<td>0.562</td>
<td>0.688</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>0.692</td>
<td>0.701</td>
<td>0.625</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.75</td>
<td>0.751</td>
<td>0.754</td>
<td>0.688</td>
<td>0.813</td>
</tr>
<tr>
<td>0.75</td>
<td>0</td>
<td>0.25</td>
<td>1.213</td>
<td>1.286</td>
<td>1.125</td>
<td>1.313</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>1.493</td>
<td>1.531</td>
<td>1.406</td>
<td>1.594</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.75</td>
<td>1.772</td>
<td>1.789</td>
<td>1.686</td>
<td>1.875</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2.049</td>
<td>2.054</td>
<td>1.968</td>
<td>2.156</td>
<td></td>
</tr>
</tbody>
</table>
Note that for the $H_2$ distribution we have computed $\alpha_1$ using equation (3.1b) with $k=2$. If $k$ is increased then the distribution moves closer to the GE-type distribution. Therefore, in reality, we are investigating the effect of $k$ (in the expression for $\alpha_1$) on the $E_2/H_2/1$. Note also that in Tables 4.1-4.5 G/P stands for Gelenbe-Pujolle formula given in [GELE 80]:

$$\langle n \rangle_D = \rho + \frac{\rho (C_a^2 + C_s^2)}{2(1 - \rho)}$$
Table 4.6 Summary of the results obtained using spectral methods

<table>
<thead>
<tr>
<th>System</th>
<th>DF*(θ)DA*(-θ) -NF*(θ)NA*(-θ)</th>
<th>Mean value formula &lt;n&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_2/H_2/1$</td>
<td>$\theta^3 - (4\lambda - \mu_1 - \mu_2)\theta^2 + [4\lambda^2 - 4\lambda (\mu_1 + \mu_2) + \mu_1 \mu_2]\theta$ $- 4\lambda \mu_1 \mu_2 (1-\rho) = 0$</td>
<td>$\frac{1+y_\text{a} + \rho^2 y_\text{s}}{1-\rho} + \frac{\lambda}{\theta_4} - 1, \theta_4 &gt; 0$</td>
</tr>
<tr>
<td>$E_2/E_2/1$</td>
<td>$\theta^3 + 4(\mu - \lambda)\theta^2 + 4(\mu^2 - 4\lambda \mu + \lambda^2)\theta$ $- 16\lambda \mu (\mu - \lambda) = 0$</td>
<td>same as $E_2/H_2/1$</td>
</tr>
<tr>
<td>$H_2/H_2/1$</td>
<td>$\theta^3 + (\mu_1 + \mu_2 - \lambda_1 - \lambda_2)\theta^2 + [\lambda_1 \lambda_2 + \mu_1 \mu_2 - (\lambda_1^2 + \lambda_2^2) (\mu_1 + \mu_2) + (\alpha_1 \mu_1 + \alpha_2 \mu_2)]\theta$ $\gamma_1 \lambda_1 \gamma_2 \lambda_2 \theta$ $\frac{\mu_1 \mu_2 \lambda_1 \lambda_2}{\lambda} (1-\rho) = 0$</td>
<td>$\frac{1+y_\text{a} + \rho^2 y_\text{s}}{1-\rho} + \frac{\lambda}{\theta_4} - \frac{\lambda(\lambda_1 + \lambda_2)}{\lambda_1 \lambda_2}, \theta_4 &gt; 0$</td>
</tr>
<tr>
<td>$E_{2}/E_{2}/1$</td>
<td>$\theta^3 + (4\lambda - \lambda_1 - \lambda_2)\theta^2 + [4\lambda^2 - 4\lambda (\lambda_1 + \lambda_2) + \lambda_1 \lambda_2]\theta$ $+ 4\mu \lambda_1 \lambda_2 (1-\rho) = 0$</td>
<td>same as $H_2/H_2/1$</td>
</tr>
</tbody>
</table>
4.3 Maximum Entropy and the G/G/1

Two major facts stem from the previous section. Firstly, the average number of customers in the system has always been obtained using an imbedded Markov process defined on the waiting time (eqn. (4.1)) without directly involving the probability distribution for the number of customers in the system. The present results give no clue as to how this distribution can be obtained from the Markov process $W_{n+1}$.

Secondly, it is obvious that the average number of customers in the system and the other performance metrics are greatly affected by the G-type model chosen for the interarrival or service time distribution. For the same values of $\rho$, $C_a^2$ and $C_s^2$ different types of distribution yield different measures of performance. These measures can be invariant to the service time distribution only in the special cases of the M/G/1 and the GE/G/1. The question is, what type of service or interarrival model gives the best estimates for the performance metrics.

In this section the maximum entropy formalism is applied in an attempt to give an answer to this question.
4.3.1 A maximum entropy model for general single queues

It is well known from queueing theory \( [\text{KLEI 75}] \) that for any single server queue with infinite capacity, if \( \rho < 1 \), then the equilibrium probability that an outside observer sees an empty system is given by

\[
p_0 = 1 - \rho
\]

It is also known that in this case the average number of customers in the system \( <n> \) is finite.

In view of this, the maximum entropy model described in section (3.1) can be generalised to any infinite capacity, single server queue. Therefore, to obtain the outside observer probability distribution \( p_n \) one maximises the entropy

\[
H(p) = \sum_{n=0}^{\infty} p_n \ln p_n
\]

subject to the constraints

\[
\sum_{n=0}^{\infty} p_n = 1
\]

\[
\sum_{n=0}^{\infty} n h(n) p_n = \rho \quad \text{where} \quad h(n) = \begin{cases} 0 & n = 0 \\ 1 & n > 0 \end{cases}
\]

and

\[
\sum_{n=0}^{\infty} n p_n = <n>
\]

which gives

\[
p_n = \begin{cases} (1-\rho) & n = 0 \\ (1-\rho)g^nx & n > 0 \end{cases}
\]  

(4.39)

where \( x \) and \( g \) can be determined from the equations.
\[
\frac{g_x}{1-x} = \frac{\rho}{1-\rho}
\]  
\[(4.40)\]

\[
\frac{g_x}{(1-x)^2} = \frac{<n>}{1-\rho}
\]  
\[(4.41)\]

(see section (3.1)).

A similar argument applies for the arriving customer distribution.

Defining \(H_n\) to be the equilibrium probability distribution that an arbitrary arriver finds \(n\) customers in the system and \(<n>\_a = \sum_{n=0}^{\infty} n \Pi_n\), then the maximum entropy arriver distribution can be obtained using the above procedure with \(p_n\) replaced by \(H_n\), \(\rho\) replaced by \(1-H_0\) and \(<n>\_a\) replaced by \(<n>\_a\). This leads to the solution:

\[
\Pi_n = \begin{cases} 
H_0 & n=0 \\
H_0 g_a x^n & n>0 
\end{cases}
\]  
\[(4.42)\]

where \(H_0\) is given by (4.8), \(g_a\) and \(x_a\) can be determined using relations similar to (4.40) and (4.41).

Now, one may ask the question, how can the maximum entropy solutions (4.39) and (4.42) be related to the steady state solution of the underlying stochastic process in the case of Markovian queues? An easy answer to this question may not be feasible at this stage. However, one can conjecture that the form of the maximum entropy solution is the same as that of the steady state solution for the \(G/GE/1\). This conjecture is based on the results obtained for the \(M/G/1\) and the \(G/M/1\) in Chapter 3 and on the properties of the \(GE\)-type distribution. Based on this we proceed to consider the \(G/GE/1\).
4.3.2 Maximum entropy and the G/GE/1

a. The outside observer distribution

Theorem (4.5)

The maximum entropy probability distribution that an outside observer sees \( n \) customers in the G/GE/1 queueing system is given by:

\[
p_n = \begin{cases} 
(1 - \rho) & n = 0 \\
(1 - \rho) g^x_n & n > 0 
\end{cases}
\]

where

\[\begin{align*}
\text{b. } x &= \frac{\mu - \theta_1 (1 + y_s)}{\mu - y_s \theta_1} \\
\text{c. } g &= \frac{\rho}{1 - \rho} - \frac{\theta_1}{\mu - \theta_1 (1 + y_s)}
\end{align*}\]

and \( \theta_1 \) is the zero of the polynomial (4.13).

Proof

For the G/GE/1, \( \langle n \rangle = \frac{\lambda}{\theta_1} - \rho y_s \)

Substituting this in (4.40) and (4.41) and solving, we get

\[
\begin{align*}
x &= \frac{\mu - \theta_1 (1 + y_s)}{\mu - y_s \theta_1} \\
g &= \frac{\rho}{1 - \rho} - \frac{\theta_1}{\mu - \theta_1 (1 + y_s)}
\end{align*}
\]

Q.E.D.
Lemma (4.1)

The value of $x$ in the solution (4.43) satisfies the following relation:

$$x = A^* \frac{\mu'(1-x)}{1-x(1-r_s)}$$

(4.44)

Proof

From (4.43b) it can be easily deduced that

$$\theta_1 = \frac{\mu'(1-x)}{1-x(1-r_s)}$$

(4.45)

since $\theta = -\theta_1$

\[.\therefore \theta = \frac{-\mu'(1-x)}{1-x(1-r_s)}\]

Making this substitution into (4.13) we get (4.44).

Q.E.D.

Note that for the G/M/1 (4.44) reduces to

$$x = A^*(\mu(1-x))$$

which is the famous secular equation for the G/M/1.

Based on Lemma (4.1) it may be concluded that the maximum entropy solution for the G/GE/1 corresponds to an imbedded Markov process at arrival instants. The properties of this process should be further investigated.
Box 4.1

For any queueing system with uncorrelated arrival and service time:

(i) the maximum entropy solution that an outside observer sees $n$
customers in the system is given by

$$p_n = \begin{cases} (1-\rho) & n=0 \\ (1-\rho)g x^n & n>0 \end{cases}$$

where

$$g = \frac{\rho(1-x)}{x(1-\rho)}$$

and

$$x = \frac{<n>-\rho}{<n>}$$

(ii) the maximum entropy solution that an arriving customer sees $n$
customers in the system

$$\Pi_0 = \begin{cases} \Pi_0 & n=0 \\ \Pi_0 g_a x^n & n>0 \end{cases}$$

where

$$g_a = \frac{(1-\Pi_0)(1-x)}{\Pi_0 x}$$

$$\Pi_0 = k = \lim_{\theta \to 0} \frac{\psi_+(\theta)}{\theta}$$

and $x$ as above.
For the G/GE/1:

\[ p_n = \begin{cases} 
(1-\rho) & \text{n=0} \\
(1-\rho)g \left[ \frac{\mu - \theta_1 (1+y_s)}{\mu - \theta_1 y_s} \right]^n & \text{n>0}
\end{cases} \]

where \( g = \frac{\rho}{1-\rho} - \frac{\theta_1}{\mu - \theta_1 (1+y_s)} \)

\[ \pi_n = \begin{cases} 
\theta_1 / \mu' & \text{n=0} \\
\frac{r_0 \theta_1}{\mu_1} \left[ \frac{\mu - \theta_1 (1+y_s)}{\mu - \theta_1 y_s} \right]^n & \text{n>0}
\end{cases} \]
The Arriver's and Completer's Distribution

Based on the form of distribution (4.42), Lemma (4.1) and the results for the M/G/1 and G/M/1 queueing systems, one can conjecture that for the arriver's distribution (4.42)

\[
x_a = x = \frac{\mu - \theta_1 (1+y_s)}{\mu - y_s \theta_1}
\]

(4.46)

Feeding this information in the constraint

\[
\sum_{n=0}^{\infty} \Pi_n = 1
\]

and knowing that \( \Pi_0 = K = \theta_1 / \mu' \), we get

\[
g_a = r_s
\]

(4.47)

Therefore, the arriver's distribution is given by

\[
\Pi_n = \begin{cases} 
\theta_1 / \mu' & \text{n=0} \\
(r_s \theta_1 / \mu')(\frac{\mu - \theta_1 (1+y_s)}{\mu - y_s \theta_1}) & \text{n>0}
\end{cases}
\]

(4.48)

Defining \( \Pi_n' \) to be the completer's distribution, then it has been shown that

\[
|KLEI 75| \quad \Pi_n' = \Pi_n
\]

(4.49)

for any G/G/1 system.
4.4 Maximum Entropy and Operational Analysis

As mentioned in Chapter 1 operational analysis has been introduced by Denning and Buzen as an alternative mathematical framework for queueing systems. In this section, we discuss how operational assumptions and laws can be incorporated in the maximum entropy formalism. First, these assumptions and laws are introduced.

4.4.1 Operational assumptions, laws and theorems

Operational laws are equations relating operational variables over a finite observation period T. Operational variables are either basic quantities which are directly measurable over the observation period or derived quantities which are obtained from the basic quantities. Operational theorems are propositions derived from operational quantities subject to testable assumptions. There are three basic testable assumptions in operational analysis:

1. The system must be flow balanced during the observation period T i.e. the number of arrivals is equal to the number of completions.
2. The device must be homogeneous i.e. the routing of jobs must be independent of local queue lengths.
3. The behaviour sequence of the queue (Fig. 4.3) during the observation period must satisfy one step transition, which implies that the queue length can change by only one job/unit time.

Adopting the same notation as Denning and Buzen, the following operational variables can be defined:

T: the length of the observation period
A: the number of arrivals during the observation period
B: the total amount of time during which the system is busy (B<T)
Fig. 4.3 Behaviour sequence of a queue for a 10-second observation period
C: the number of completions occurring during the observation period.

Y₀: the overall arrival rate = A/T

Yₙ: arrival rate when n(t) = n

N: the total number of customers in the system during the observation period.

X: the output rate = C/T

U: utilisation = B/T

S: the overall mean service time per completed job = B/C

S(n): the mean service time per completed job when n(t) = n.

The assumption of job flow balance would imply that:

A = C \quad (4.50)

Similarly, the homogeneity assumption implies that

Y(n) = Y \quad \text{and} \quad S(n) = S \quad (4.51)

for all n

Based on these assumptions and the one step transition assumption, the following operational queue recursions are satisfied:

\[ p(n) = SY(n-1) \quad n=1, \ldots, N \quad (4.52) \]

\[ p_A(n) = SY(n-1) \quad n=1, \ldots, N-1 \quad (4.53) \]

where \( p(n) \) is the operational outside observer distribution

and \( p_A(n) \) is the arriving customer operational distribution.
hold even when the observation period $T \to \infty$.

Defining $\langle n \rangle$ to be

$$\langle n \rangle = \lim_{T \to \infty} \langle n \rangle_T$$

it may be a reasonable approximation to replace $\langle n \rangle$ by its Markovian counterpart if $U < 1$. In this way maximum entropy provides a useful link between Markovian and operational analysis. The advantages of this link will be more obvious in Chapter 6.

Moreover the maximum entropy recursion may be very useful in verifying the operational recursion. The maximum entropy recursion will reduce to the operational recursion if $x_T = SY$ and $g_T = 1$. If that does not happen, then this implies that some information is missing from the model.

The above arguments similarly apply for the arriver's distribution.
4.5 Flow in the G/GE/1 and the GE/GE/1 Systems

Following the discussion of the previous sections, it can be realised that the G/GE/1 and the GE/GE/1 are very important for the analysis of general queueing networks. This is because:

a. The steady state solution for the number of customers in the system is expected to be closest, if not identical, to the corresponding maximum entropy solution of the system.

b. It is very easy to compute many of the flow equations for these systems (interdeparture and interinput distributions).

c. Since the GE-type distribution is very close to the pure exponential, this distribution is expected to minimise the correlation in the flow.

d. The structure of these systems may give a possible explanation for the robustness of exponential queueing networks.

For these reasons, it appears worthwhile to consider the flow equations for the above systems. In section (4.6.1) the idle time and the interdeparture time distributions are identified for both systems. In Section 4.6.2 the interinput distribution for the GE/GE/1 with Bernoulli feedback is derived.

4.6.1 The idle time distribution and the departure process for the G/GE/1

It is known that the interdeparture time distribution is closely related to the idle time and service time distribution. Defining $D^*(\theta)$ to be the Laplace transform of the interdeparture time distribution, then
Marshall \cite{MARS68} has shown that:

\[ D* (\theta) = \Pi_0 (I* (\theta) F* (\theta)) + (1-\Pi_0) F* (\theta) \tag{4.58} \]

which holds for all variations of the G/G/1.

The idle time distribution for the G/GE/1 can be deduced from (4.9) and (4.19) as

\[ I* (\theta) = 1 + \theta / \psi_- (\theta) \]
\[ = \frac{\theta_1 - \mu' (1-r_s) \theta + \mu' I* (\theta)}{\theta_1 + \theta} \tag{4.59} \]

In this section (4.58) and (4.59) are utilised to obtain some useful results for the departure process: the analysis is started by the following theorem:

\textbf{Theorem 4.6}

The first moment of the interdeparture time distribution is given by

\[ \overline{D} = E(t) = 1 / \lambda \tag{4.60} \]

The second moment of the departure process is

\[ \overline{D^2} = E(t^2) = \overline{f^2} + \Pi_0 \{ \overline{I^2} + 2 \overline{I} \overline{f} \} \tag{4.61} \]

where \( \overline{I} \) is the first moment of the idle time distribution and is given by:

\[ \overline{I} = \frac{\mu' a - r_s}{\theta_1} \tag{4.62} \]

\( \overline{I}^2 \) is the second moment of idle time distribution and is given by:

\[ \overline{I}^2 = 2 \{ \theta_1 (1-r_s) \overline{a} + \mu' \theta_1 \overline{a}^2 / 2 + r_s - \mu' \overline{a} \} / \theta_1^2 \tag{4.63} \]
\( \bar{a} \) is the first moment of the interarrival time distribution given by

\[
\bar{a} = 1/\lambda
\]

\( \bar{a}^2 \) is the second moment of the interarrival time distribution given by

\[
\bar{a}^2 = \left(\frac{C_a^2 + 1}{\lambda^2}\right)
\]

and \( \bar{f} \) is the first moment of the service time distribution and is given by

\[
\bar{f} = 1/\mu
\]

**Proof**

The first moment can be obtained from the Laplace transform using the following relation

\[
\bar{D} = -\frac{dD^*(\theta)}{d\theta} \bigg|_{\theta=0}
\]

(see [KLEI 75a]{})

from (4.58)

\[
\frac{dD^*(\theta)}{d\theta} = I_0 \{ I*(\theta) \frac{dF^*(\theta)}{d\theta} + F*(\theta) \frac{dI^*(\theta)}{d\theta} \} + (1-I_0) \frac{dF^*(\theta)}{d\theta}
\]

since

\[
\bar{I} = -dI*(\theta)/d\theta \bigg|_{\theta=0}, \quad \bar{f} = -dF*(\theta)/d\theta \bigg|_{\theta=0}, \quad I*(\theta) = F*(\theta) = 1
\]

we immediately get

\[
\bar{D} = I_0 \bar{I} + \bar{f}
\]

Similarly

\[
\bar{D}^2 = \frac{dD^*(\theta)^2}{d\theta^2} \bigg|_{\theta=0}
\]
which, with a little calculus and algebra, gives

\[
\frac{D^2}{I} = \frac{1}{I_0} + \frac{I^2}{I_0^2} + 2 \frac{I}{I_0} \left( I + 2 \frac{I}{I_0} \right)
\]

The moments of the idle time distribution can be obtained from the fact that

\[
I^*(\theta) = \int_0^\infty e^{\theta t} dI(t) \quad \text{and} \quad A^*(\theta) = \int_0^\infty e^{\theta t} dA(t)
\]

which gives

\[
\bar{I} = \frac{dI^*(\theta)}{d\theta} \bigg|_{\theta=0} \quad \text{and} \quad \bar{I}^2 = \frac{dI^*(\theta)}{d\theta} \bigg|_{\theta=0}
\]

Performing these operations on (4.59) above we get (4.62) and (4.63) respectively. Substituting (4.62) and (4.64), we get \(\bar{D} = 1/\lambda\).

Q.E.D.

Theorem (4.7)

For the GE/GE/1 system:

\[
a(t) = (1-r_a)u_o(t) + r_a \lambda'e^{-\lambda't}
\]

a. The idle time pdf is given by

\[
I(t) = (1-L)u_o(t) + L\lambda'e^{-\lambda't}
\]

(4.65)

where

\[
L = \frac{r_a + r_s - r_a r_s}{r_a}
\]

(4.66)
b. The interdeparture time pdf is given by

\[ D(t) = E_1 u_o(t) + E_2 r e^{-\mu t} + E_3 L P e^{-\lambda t} \]  \hspace{1cm} (4.67)

where

\[ E_1 = (1-r_s)(1-L \Pi o) \]  \hspace{1cm} (4.68)

\[ E_2 = \frac{\mu'(1-L \Pi o) - \lambda'}{\mu' - \lambda'} \]  \hspace{1cm} (4.69)

and

\[ E_3 = \frac{\mu' - (1-r_s)\lambda'}{\mu' - \lambda'} \]  \hspace{1cm} (4.70)

**Proof**

In this case \( \psi_\infty(\theta) = \frac{(\lambda' - \theta)}{(r_a + r_s - r_a r_s)} \).

Defining \( L = r_a + r_s - r_a r_s \) and using (4.59) we get

\[ I^*(\theta) = 1 + \frac{L \theta}{\lambda' - \theta} \]

which, on manipulation, gives

\[ I^*(\theta) = (1-L) + \frac{L \lambda'}{\lambda' + \theta} \]  \hspace{1cm} (4.71)

which on inversion gives (4.65).

Substituting (4.71) in (4.58) and manipulating, we get

\[ D^*(\theta) = (1-r_s)(1-L \Pi o) + \frac{r_s \mu'}{\mu' + \theta} (1-L \Pi o - \frac{L \Pi o \lambda'}{\mu' - \lambda'}) + \frac{L \Pi o \lambda'}{\lambda' + \theta} (1-r_s + \frac{r_s \mu'}{\mu' + \lambda'}) \]  \hspace{1cm} (4.72)

which on re-arrangement and inversion gives (4.67).

Q.E.D.
For the G/GE/1 system

\[
\bar{D} = 1/\lambda , \quad \bar{D}^2 = \bar{f}^2 + \Pi_0 \{ \bar{I}^2 + 2\bar{I} \bar{f} \}
\]

\[
\bar{I} = \frac{\mu' \bar{a} - \bar{r}_s}{\bar{a}} \quad \bar{a} = 1/\lambda \quad , \bar{f} = 1/\mu
\]

\[
\bar{I}^2 = 2\{ \theta_1 (1-\bar{r}_s) \bar{a} + \mu' \theta_1 \bar{a}^2 /2 + \bar{r}_s - \mu' \bar{a} \} / \theta_1^2
\]

\[
\bar{r}_s = 2/(C_s^2 + 1)
\]

\[
I^*(\theta) = 1 + \theta / \psi_-(\theta)
\]

\[
D^*(\theta) = \Pi_0 (I^*(\theta) F^*(\theta)) + (1-\Pi_0) F^*(\theta)
\]

\(\theta_1\) is the zero of the polynomial

\[
\mu' \bar{a} - \{ \mu' (1-\bar{r}_s \theta) \} A^*(-\theta)
\]
For the GE/GE/1 system, 

\[ i(t) = (1-L)u_o(t) + L\lambda^e e^{-\lambda't} , \text{ with } L = r_a + r_s - r_a r_s \]

\[ r_a = 2/(c_a^2 + 1) \quad r_s = 2/(c_s^2 + 1) \]

\[ D(t) = E_1 u_o(t) + E_2 r_s \mu^e e^{-\mu't} + E_3 L\lambda^e e^{-\lambda't} \]

with

\[ E_1 = (1-r_s)(1-L\Pi_o), \quad E_2 = (\mu'(1-L\Pi_o) - \mu')/(\mu' - \lambda') \]

\[ E_3 = (\mu' - (1-r_s)\lambda')/(\mu' - \lambda') \]
4.5.2 The GE/GE/1 with Bernoulli feedback

If, for the GE/GE/1 considered above, customers are allowed to rejoin the queue after finishing service with probability $q_1$ or totally leave the system with probability $q_2$, then we have a GE/GE/1 with Bernoulli feedback. It has been shown [BURK 76], [DISN 77] that this sort of feedback greatly changes the features of the flow in the system e.g. Burke [BURK 76] showed that for the M/M/1 with Bernoulli feedback, the input process is no longer Poisson. Disney and McNickel [DISN 76] studied the M/G/1 with Bernoulli feedback and concluded that the input process may not even be a renewal process.

In this section some of the results obtained for the M/M/1 and the M/G/1 are generalised for the maximum entropy GE/GE/1. The system considered is shown in Fig. 4.5. Let $A'(t)$ be the PDF of the inter-input times (point 2 on the figure) with Laplace transform $A^*(\theta)$ and $G(t)$ be the probability that the only customer to arrive between $t$ and $t+\Delta t$ to be a feedback customer. If the arrival time PDF is taken to be

$$A'(t) = 1 - r_a e^{-\lambda't}$$

(4.73)

then, following Burke [BURK 76], Pujolle and Soula [PUJO 79], we get

$$1 - A'(t) = r_a e^{-\lambda't} [1 - G(t)]$$

(4.74)

Using Laplace transform this becomes

$$\frac{1}{\theta} - \frac{A^*(\theta)}{\theta} = r_a \left\{ \frac{1}{\lambda' + \theta} - \frac{1}{\lambda' + \theta} G^*(\lambda' + \theta) \right\}$$

which gives

$$A^*(\theta) = (1 - r_a) + \frac{r_a \lambda'}{\lambda' + \theta} \left\{ 1 - G^*(\lambda' + \theta) \right\} + r_a G(\lambda' + \theta)$$

(4.75)
Fig. 4.4 A queue with feedback
To evaluate $A^*(\theta)$, an expression for $G^*(\lambda+\theta)$ should be obtained.

Pujolle and Soula [PUJO 79] have shown that $G^*(\theta)$ is given by

$$G^*(\theta) = \sum_{i=0}^{\infty} \Pi(i) q_1 q_2^{i+1} \left( F^*(\theta) \right)^i$$

where

$$\Pi(i) = \sum_{j=i}^{\infty} \Pi(j)$$

For the GE/GE/1 $\Pi(i)$ is given by (4.42) from which

$$\Pi(i) = \begin{cases} 1 & \text{if } i=0 \\ (1-\Pi_0) x^{i-1} & \text{for } i > 0 \end{cases}$$

where

$$x = (\rho+y_a+py_s)/(1+y_a+py_s), \quad \rho = \lambda/q_2 \mu$$

$F^*(\theta)$ is the Laplace transform of the service time distribution given by (4.11). On manipulation (4.78) gives

$$G^*(\theta) = q_1 L(1-r_s) \left[ 1-q_2(x-\sigma)(1-r_s) \right] + \frac{PE}{LB+\theta} \left( 1-q_2(x-\sigma)(1-r_s) \right)$$

$$- \frac{r_s \mu'}{\mu' + \theta} (Lq_2 q_2(x-\sigma)(1-r_s)) - \frac{q_1 q_2 r_s (x-\rho)E}{(\mu'+\theta)(LB+\theta)}$$

where

$$\sigma = 1-\Pi_0, \quad L = (1-q_2 x (1-r_s))^{-1}, \quad B = \mu'(1-q_2 x)$$

and

$$E = L(\mu'-LB(1-r_s))$$

Substituting this in (4.77) and manipulating, we get

$$A^*(\theta) = (1-\beta \sigma L T_2) + \frac{\sigma q_1}{1-T_1-T_2-T_3} \frac{\lambda'}{\lambda' + \theta}$$

$$+ \frac{\sigma a T_2}{LB+\lambda' + \theta} + \frac{\sigma a T_3 (\mu' + \lambda')}{\mu' + \lambda' + \theta}$$

where

$$T_2 = q_1 (1-\gamma_3 q_1 q_2 \sigma (1-\gamma_3)), \quad T_3 = q_1 L \sigma \mu' \sigma / B x$$

and

$$T = T_2$$

(4.82)
Therefore, the interinput distribution is a mixture of exponentials.

The idle time distribution for this system can still be determined using (4.9) with $A^*(\theta) = A^* (\theta)$. If this distribution is obtained then (4.58) can be used to obtain the interoutput time distribution with $\rho = \lambda / q_2 \mu$.

4.6 Summary

In the first part of this chapter, the spectral methods have been extended to give mean value formulae for most of the known types of G/G/1 queueing systems. These results are summarised in Table 4.6. The analysis required the isolation of positive or negative zeros for a polynomial in $\theta$ which has been effectively done by the use of Rouche's theorem. The numerical results obtained (Tables 4.1-4.5) illustrated how the average number of customers in the system can be affected by our assumptions about the service and interarrival time distributions.

To avoid arbitrary assumptions about these distributions and to obtain a suitable probability distribution for the number of customers in the system, maximum entropy has been suggested as a criterion and a method of solution. A maximum entropy model was developed for general queues. This model gives all three types of the probability distribution for the number of customers in the system based on the expected number of customers in the system and the percentage of time the server is busy. These results are shown in Box 4.1. Following this it has been explained that for the G/GE/1 the maximum entropy solution can be identical or very close to the steady state solution of the underlying stochastic process. For this reason, the system can be of special importance in the analysis of general queueing networks. Because of this, a section has been devoted
to study the flow processes for this system. These include the departure
process, the idle time distribution and the interrupt distribution for
the GE/GE/1 with feedback. The maximum entropy solution for the G/GE/1
are summarised in Box 4.2 and the flow equations are given in Box 4.3.

Before that the maximum entropy model has been compared with
operational models. It has been shown that the maximum entropy model
leads to very useful queue recursions similar to those of operational
analysis but without the need for operational assumptions. These
recursions imply that the maximum entropy solution for the G/G/1 satisfies
local balance, which is an important result for the analysis of general
queueing networks.
CHAPTER V
TWO-SERVER TANDEM AND CYCLIC QUEUES WITH
NON-EXPOENTIAL SERVICE TIMES

5.1 Introduction

In many practical applications one encounters queueing systems in which the service facilities are arranged in 'tandem'. A newly arriving customer queues for the first facility and after completing service it either leaves the system or joins the queue for the next facility in the series. (Fig. 5.1). This type of queue is known in the literature as tandem or series queues. If the output of the last queue in the series is partially or totally connected to the first queue, then the system becomes a cyclic queueing model which can be open (Fig. 5.2a) or closed (Fig. 5.2b). These systems represent the simplest types of queueing networks.

Tandem queues were first introduced by Taylor and Jackson [TAYL 54], then Jackson [JACK 54,56]. Taylor and Jackson also introduced the concept of cyclic queues, but the first application of these models was due to Koenigsberg [KOEN 58] where he applied the models in the area of conventional mining. Application to information systems started with Kleinrock [KLEI 66] in the analysis of sequential processing machines. Mitrani [MITR 72] suggested a two-server cyclic queueing model to analyse a multiprogramming computer system. In all these models an exponential distribution was assumed for the service time at all centres. Mitrani showed that his model can be analysed using an equivalent M/M/1/N system. Reiser and Kobyashi [REIS 74] showed that if an arbitrary service time is introduced at one of the centres in the two-server cyclic model,
Fig. 5.1 A tandem queueing system

Fig. 5.2a An open cyclic queueing model

Fig. 5.2b A closed cyclic queueing model
then the system can be analysed using a single equivalent M/G/1/N system. This equivalence principle is in fact a special case of the 'dominance principle' introduced earlier by Friedman [FRIE 65] for general tandem queues. The dominance principle implies that, if certain conditions are satisfied, then the analysis of a tandem queueing system can be reduced to the analysis of an equivalent system with fewer servers, irrespective of the service time distributions at the individual servers. Koenigsberg [KOEN 82] noted that the dominance principle applies as well to closed cyclic queueing models.

In this chapter the analysis of two-server tandem and cyclic queueing models with general service times at both centres, is considered. Although a direct maximum entropy model can be applied here, the systems are analysed using Markovian queueing theory while taking advantage of the G/G/1 results obtained in the previous chapter. Application of the maximum entropy formalism is left for the next chapter.

Since the G/G/1 with finite waiting room is important to the analysis of closed two-server cyclic models, this system is considered in section 5.2. Analysis of tandem and cyclic queues is carried out in section 5.3 accompanied with various examples and comparisons. The correlation in the departure process and its implication on the analysis are discussed in section 5.4. Concluding remarks are given in section 5.5. The motivation is to illustrate the complexity that arises in the stochastic analysis of general queueing networks, and to solve some of the problems relating to these models. The results obtained will also be used in the aggregation of general queueing networks to be considered later.
5.2 The G/G/1 with Finite Waiting Room

In this section the G/G/1 system with finite waiting positions, N-1 is considered. The analysis is based on the approach used to analyse the M/G/1/N |KOBA 78 |, |COOP 81 | and some few conjectures.

When the number of waiting positions is finite, then a customer will depart either after he has completed service or if he arrives when all positions are occupied i.e. without taking service |COOP 81 |. In this sense there are N+1 states: (0, 1, ..., N), and the system is shortly referred to as the G/G/1/N.

Defining pN(n) to be the maximum entropy equilibrium probability that a random observer sees n customers in the system, and retaining the form of constraints for the infinite capacity G/G/1, one can conjecture that:

\[ p_N(n) = C_p \quad \text{for} \quad 0 \leq n \leq N-1 \]  \hspace{1cm} (5.1)

where \( p_n \) is the equilibrium maximum entropy probability distribution that a random observer sees n customers in the infinite room G/G/1, and is given by (4.39). This conjecture is based on the results obtained for the M/G/1/N and the properties of the maximum entropy G/G/1.

Furthermore the normalisation condition implies that

\[ \sum_{n=0}^{N} p_N(n) = 1 \]  \hspace{1cm} (5.2)

The effective server utilisation can be defined as

\[ \rho' = 1 - p_N(0) \]  \hspace{1cm} (5.3)
For the G/GE/1 with finite waiting room, the maximum entropy solution:

\[
p_N(n) = \begin{cases} 
  p_N(O)gx^n & \text{for } n = 0 \\
  \frac{\rho - (1-p_N(O))}{\rho} & \text{for } 1 \leq n < N-1 \\
  \frac{1}{\rho} & \text{for } n = N
\end{cases}
\]

where \( x = \frac{\mu' - \theta_1}{\mu' -(1-r_s)\theta_1} \)

\[
g = \frac{\rho(1-x)}{x(1-\rho)}
\]

and \( \theta_1 \) is the zero of the polynomial

\[
(\mu' + \theta - ((1-r_s)\theta + \mu')A(-\theta) = 0
\]
Since $p_N(0)$ is the fraction of customers who leave without service, the effective service utilisation can also be expressed as

$$\rho' = \rho(1 - p(N)) \quad (5.4)$$

where $\rho = \lambda/\mu$

From (5.3) and (5.4) it can be easily deduced that:

$$p(N) = \frac{\rho - (1 - p_N(0))}{\rho} \quad (5.5)$$

Using (5.1), (5.2) and (5.5), $C$ can be obtained as:

$$C = \frac{1}{(1+\rho)p_o + \rho \sum_{n=1}^{N-1} p_n}$$

$$= (1-x) p_o ((1+\rho)(1-x) + \rho g x (1-x^{N-1})) \quad (5.6)$$

$$p_N(n) = \begin{cases} p_N(0) g x^n & 1 \leq n \leq N-1 \\ \frac{\rho - (1-p_N(0))}{\rho} & n=N \end{cases} \quad (5.7)$$

from which the average number of customers in the system is

$$<n> = \sum_{n=1}^{N} n p_N(n)$$
Other performance metrics can be obtained using Little's formula

\[
\langle n \rangle = \lambda_n T = \lambda_n (W + \frac{1}{\mu})
\]  

(5.9)

where \( \lambda_n = \lambda (1 - p_N(n)) \)

(5.10)

and \( T \) is the response time in the \( G/C/1/N \).

Note that the imbedded distribution at departure instants can similarly be obtained using the relation

\[
\Pi'_N(n) = C' \Pi'_n \quad 0 \leq n \leq N-1
\]

(5.11)

and the condition

\[
\sum_{n=0}^{N-1} \Pi'_N(n) = 1
\]

(5.12)

The \( N \)th state is not included since there are only \( N \) states from the departing customer's point of view. (5.11) and (5.12) gives \( C' \) as

\[
C' = 1/(\Pi'_0 (1-x) + r_s x (1-x^N))
\]

(5.13)

In the case of the arriving customer distribution the \((N)\)th state is included and the probabilities \( \Pi'_N(n) \) can be determined if relations similar to (5.3) and (5.4) can be deduced.
Fig. 5.3 A two-station tandem queue, example 5.3.1 (i-iii)

Fig. 5.4 A two-station closed cyclic model, example 5.3.2(i,ii)
For the G/GE/1

\[ x = \frac{(\mu' - \theta_1)}{(\mu' - (1 - r_s) \theta_1)} \]

and \( \theta_1 \) is the zero of the equation

\[ (\mu' + \theta) - ((1 - r_s) \theta + \mu')A(-\theta) \]

and \( 0 < r_s < 2 \) (see Chapter 4).

5.3 Analysis of Tandem and Cyclic Queues

In this section two basic models and their variations are considered. The first is a general two-station tandem queueing system with an infinite number of customers. The second is a two-station closed cyclic queueing model with a finite number of customers \( N \). The analysis is carried out using both the G/GE/1 and the GE/GE/1 defined in the previous chapter. The two systems are compared and some conclusions are drawn.

5.3.1 A general two-station tandem queueing system

Consider the model shown in Fig. 5.3. For this model customers arriving from outside the system join the queue for station 1. After finishing service at this station they join the queue for station 2 and after that they totally leave the system. Each queue has an infinite waiting room and service is provided on a FCFS basis.

Assume that information about the flow at each queue is given in the shape of the respective mean rates and coefficients of variations which are not necessarily equal to one. Then, following the results
obtained in Chapter 3, we will assume that the inter-arrival time distribution for station 1 will be of the form:

\[ a_1(t) = (1-r_{al})u_0(t) + r_{al}\lambda_1't \]

where \( r_{al} = 2/(C_{al}^2 + 1) \) and \( \lambda_1' = r_{al}\lambda_1 \)

and the service time distribution will be of the form:

\[ f_1(t) = (1-r_{sl})u_0(t) + r_{sl}\mu_1'e^{-\mu_1't} \]

where \( r_{sl} = 2/(C_{sl}^2 + 1) \) and \( \mu_1' = r_{sl}\mu_1 \)

The justification for these assumptions is that the steady state solutions corresponding to these distributions is identical or closest to the maximum entropy solution (see the previous two chapters).

However, for the second station the available information shows that the inter-input distribution \( a_2'(t) \) is given by

\[ a_2'(t) = d_1(t) \]

where \( d_1(t) \) is the inter-departure time distribution for the first station. Therefore the Laplace transform of \( a_2(t) \) is given by (using (4.58)):

\[ \mathcal{L}\{a_2'(t)\}(\theta) = \mathcal{L}\{d_1(t)\}(\theta) = \pi_{01}(I_1*\theta)F_1*\theta + (1-\pi_{01})F_1*\theta \]

The service time distribution for station 2 is again given by:

\[ f_2(t) = (1-r_{s2})u_0(t) + r_{s2}\mu_2'e^{-\mu_2't} \]

where \( r_{s2} = 2/(C_{s2}^2 + 1) \) and \( \mu_2' = r_{s2}\mu \)
Since station 1 is not affected by station 2, it can be analysed as a $GE/GE/1$ queueing system.

To analyse the second station one should use the results for the $G/GE/1$. In this case the zero of equation (4.13) written as

$$\mu_2' + \theta - ((1-r_{s2})\theta + \mu_2')A_2'(-\theta) = 0$$

(5.18)

should be obtained. Or, equivalently, the value of $x_2$ can be obtained using Lemma (4.1') written as

$$x_2 = \frac{\mu_2'(1-x_2)}{A_2'(1-x_2(1-r_{s2}))}$$

(5.19)

The zero of (5.18) will then be

$$\theta_1' = \frac{\mu_2'(1-x_2)}{1-x_2(1-r_{s2})}$$

(5.20)

The equilibrium probabilities and the various performance metrics can then be obtained using the equations given in section 4.5.1.

For this model the analysis appeared to be so simple because there are only two stations. If, for example, a third station is added, then the zero of (5.18) or the solution of (5.19) will be more difficult to obtain. The complexity grows further and further with the addition of new stations. In such cases it may be advantageous to approximate the inter-input distribution for station $i$ by

$$a_i'(t) = (1-r_{d,i-1})u_0(t) + \lambda' r_{d,i-1}e^{-\lambda_i't}$$

(5.21)

where
Table 5.1 The Average Number of Customers and the Mean Waiting Time for Station 2 in the Tandem Queue of Example 5.3.1(i)

\[ \rho_1 = 0.6, \quad \rho_2 = 0.75 \]

<table>
<thead>
<tr>
<th>C&lt;sub&gt;2&lt;/sub&gt;</th>
<th>C&lt;sub&gt;2&lt;/sub&gt;</th>
<th>C&lt;sub&gt;2&lt;/sub&gt;</th>
<th>θ&lt;sub&gt;1(2)&lt;/sub&gt;</th>
<th>&lt;n&gt;_2 using</th>
<th>&lt;n&gt;_2 using</th>
<th>%</th>
<th>W&lt;sub&gt;2&lt;/sub&gt; using</th>
<th>W&lt;sub&gt;2&lt;/sub&gt; using</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>al</td>
<td>sl</td>
<td>dl</td>
<td>θ&lt;sub&gt;1(2)&lt;/sub&gt;</td>
<td>the G/GE/1</td>
<td>the GE/GE/1</td>
<td>difference</td>
<td>the G/GE/1</td>
<td>the GE/GE/1</td>
<td>difference</td>
</tr>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>0.619</td>
<td>0.5</td>
<td>0.438</td>
<td>1.898</td>
<td>1.868</td>
<td>1.6</td>
<td>1.53112631</td>
<td>1.49</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.0</td>
<td>0.306</td>
<td>2.453</td>
<td>2.430</td>
<td>0.92</td>
<td>2.27039</td>
<td>2.24</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2.0</td>
<td>0.190</td>
<td>3.570</td>
<td>3.555</td>
<td>0.418</td>
<td>3.759931</td>
<td>3.74</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4.0</td>
<td>0.108</td>
<td>5.814</td>
<td>5.805</td>
<td>0.15</td>
<td>6.751789</td>
<td>6.74</td>
</tr>
<tr>
<td>1.0</td>
<td>0.8</td>
<td></td>
<td>0.5</td>
<td>0.374</td>
<td>2.192</td>
<td>2.138</td>
<td>2.5</td>
<td>1.92332</td>
<td>1.85</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.0</td>
<td>0.273</td>
<td>2.742</td>
<td>2.700</td>
<td>1.54</td>
<td>2.656385</td>
<td>2.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2.0</td>
<td>0.177</td>
<td>3.854</td>
<td>3.825</td>
<td>0.74</td>
<td>4.13847</td>
<td>4.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4.0</td>
<td>0.104</td>
<td>6.0926</td>
<td>6.075</td>
<td>0.28</td>
<td>7.123475</td>
<td>7.1</td>
</tr>
<tr>
<td>2</td>
<td>1.16</td>
<td></td>
<td>0.5</td>
<td>0.294</td>
<td>2.7347</td>
<td>2.6775</td>
<td>2.09</td>
<td>2.646</td>
<td>2.57</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1.0</td>
<td>0.228</td>
<td>3.286</td>
<td>3.24</td>
<td>1.41</td>
<td>3.382</td>
<td>3.32</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2.0</td>
<td>0.157</td>
<td>4.398</td>
<td>4.365</td>
<td>0.76</td>
<td>4.864</td>
<td>4.82</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4.0</td>
<td>0.097</td>
<td>6.636</td>
<td>6.615</td>
<td>0.32</td>
<td>7.849</td>
<td>7.82</td>
</tr>
</tbody>
</table>
The average number of customers and the mean waiting time for station 2 in the tandem queue of example 5.3.1(ii)

<table>
<thead>
<tr>
<th>$C_{al}^2$</th>
<th>$C_{s1}^2$</th>
<th>$C_{d1}^2$</th>
<th>$C_{s2}^2$</th>
<th>$\theta_1(2)$</th>
<th>$&lt;n&gt;_2$ using the G/GE/1</th>
<th>$&lt;n&gt;_2$ ' using the GE/GE/1</th>
<th>% difference</th>
<th>$W_2$ using the G/GE/1</th>
<th>$W_2'$ using the GE/GE/1</th>
<th>% difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>0.625</td>
<td>0.5</td>
<td>0.8436</td>
<td>0.7177</td>
<td>0.6875</td>
<td>4.2</td>
<td>0.4354</td>
<td>0.375</td>
<td>13.9</td>
</tr>
<tr>
<td>0.5</td>
<td>1.0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5351</td>
<td>0.8844</td>
<td>0.8625</td>
<td>2.5</td>
<td>0.7689</td>
<td>0.725</td>
<td>5.7</td>
</tr>
<tr>
<td>0.5</td>
<td>2.0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.3762</td>
<td>1.0792</td>
<td>1.0625</td>
<td>1.5</td>
<td>1.1583</td>
<td>1.125</td>
<td>2.9</td>
</tr>
<tr>
<td>0.5</td>
<td>4.0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.2153</td>
<td>1.5729</td>
<td>1.5625</td>
<td>0.66</td>
<td>2.148</td>
<td>2.125</td>
<td>0.97</td>
</tr>
<tr>
<td>1.0</td>
<td>0.75</td>
<td>0.5</td>
<td>0.5</td>
<td>0.7464</td>
<td>0.7949</td>
<td>0.75</td>
<td>5.6</td>
<td>0.5897</td>
<td>0.5</td>
<td>15.2</td>
</tr>
<tr>
<td>1.0</td>
<td>1.0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.4956</td>
<td>0.9589</td>
<td>0.925</td>
<td>3.5</td>
<td>0.9177</td>
<td>0.85</td>
<td>7.38</td>
</tr>
<tr>
<td>1.0</td>
<td>2.0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.3568</td>
<td>1.1514</td>
<td>1.125</td>
<td>2.3</td>
<td>1.3028</td>
<td>1.25</td>
<td>4</td>
</tr>
<tr>
<td>1.0</td>
<td>4.0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.2090</td>
<td>1.6419</td>
<td>1.625</td>
<td>1</td>
<td>2.2839</td>
<td>2.25</td>
<td>1.5</td>
</tr>
<tr>
<td>2.0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>0.6265</td>
<td>0.9231</td>
<td>0.875</td>
<td>5.21</td>
<td>0.8463</td>
<td>0.75</td>
<td>11.37</td>
</tr>
<tr>
<td>2.0</td>
<td>1.0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.4393</td>
<td>1.0882</td>
<td>1.0500</td>
<td>3.5</td>
<td>1.1763</td>
<td>1.1</td>
<td>6.5</td>
</tr>
<tr>
<td>2.0</td>
<td>2.0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.3266</td>
<td>1.2808</td>
<td>1.25</td>
<td>2.4</td>
<td>1.5616</td>
<td>1.500</td>
<td>3.9</td>
</tr>
<tr>
<td>2.0</td>
<td>4.0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.1984</td>
<td>1.7707</td>
<td>1.75</td>
<td>1.2</td>
<td>2.5414</td>
<td>2.5</td>
<td>1.6</td>
</tr>
</tbody>
</table>

$p_1 = 0.5$, $p_2 = 0.5$
Table 5.3 The average number of customers and the mean waiting time for station 2 in the tandem queue of example 5.3.1(iii)

\[ \rho_1 = 0.3, \quad \rho_2 = 0.3 \]

<table>
<thead>
<tr>
<th>( c_{sl}^2 )</th>
<th>( C_{dl}^2 )</th>
<th>( C_{s1}^2 )</th>
<th>( C_{s2}^2 )</th>
<th>( \theta(1,2) )</th>
<th>( \langle n \rangle_2 ) using ( G/E/1 )</th>
<th>( \langle n \rangle_2' ) using ( G/E/1 )</th>
<th>( % ) Difference</th>
<th>( W_2 ) using ( G/E/1 )</th>
<th>( W_2' ) using ( G/E/1 )</th>
<th>( % ) Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.5</td>
<td>0.605</td>
<td>0.5</td>
<td>1.126</td>
<td>0.3414</td>
<td>0.3118</td>
<td>8.7</td>
<td>0.138</td>
<td>0.0393</td>
<td>71.53</td>
</tr>
<tr>
<td>1.0</td>
<td>0.7333</td>
<td>0.3791</td>
<td>0.3568</td>
<td>5.9</td>
<td>0.2638</td>
<td>0.1893</td>
<td>28.23</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.0</td>
<td>0.5212</td>
<td>0.4256</td>
<td>0.4082</td>
<td>4.1</td>
<td>0.4186</td>
<td>0.3607</td>
<td>13.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.0</td>
<td>0.3006</td>
<td>0.5479</td>
<td>0.5368</td>
<td>2.0</td>
<td>0.8263</td>
<td>0.7892</td>
<td>4.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.65</td>
<td>0.5097</td>
<td>0.4386</td>
<td>0.4176</td>
<td>4.72</td>
<td>0.462</td>
<td>0.3929</td>
<td>15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.74</td>
<td>0.297</td>
<td>0.5601</td>
<td>0.5464</td>
<td>2.43</td>
<td>0.867</td>
<td>0.8214</td>
<td>5.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>1.0</td>
<td>0.375</td>
<td>0.3407</td>
<td>9.14</td>
<td>0.2500</td>
<td>0.1357</td>
<td>45.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>0.6762</td>
<td>0.4136</td>
<td>0.3857</td>
<td>6.8</td>
<td>0.3788</td>
<td>0.2857</td>
<td>24.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.0</td>
<td>0.4917</td>
<td>0.4601</td>
<td>0.4371</td>
<td>5</td>
<td>0.5337</td>
<td>0.4571</td>
<td>14.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.0</td>
<td>0.2908</td>
<td>0.5816</td>
<td>0.5657</td>
<td>2.7</td>
<td>0.9226</td>
<td>0.8857</td>
<td>5.6</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
\[ r_{d,i-1} = \frac{2}{(C_{d,i-1}^2 + 1)} \]  

(5.22)

and \( C_{d,i-1}^2 \) is the squared coefficient of variation of the departure process from station \( i-1 \), which is easily obtainable from Theorem (4.6), no matter how complex are the functions involved. The analysis is then carried out using the GE/GE/1 results. To evaluate this approximation three numerical examples are considered.

(i) Consider the case where \( \rho_1 = 0.6, \rho_2 = 0.75 \) and \( C_{s1}^2 = 0.5 \). In Table 5.1 the average number of customers and the mean waiting time are computed for different combinations of \( C_{s1}^2 \) and \( C_{s2}^2 \), the squared coefficients of variation at station 1 and station 2 respectively. The results are first obtained using the G/GE/1 model (\( \langle n >_2, W_2 \rangle \)) and then the GE/GE/1 model (\( \langle n >'_2, W'_2 \rangle \)) with (5.21) representing the inter-input distribution. The table shows that the maximum percentage deviation of \( \langle n >'_2 \) from \( \langle n >_2 \) is always less than 3% and the maximum deviation of \( W'_2 \) from \( W_2 \) is less than 4%. It is also observed that this deviation decreases when \( C_{s2}^2 \) increases.

(ii) If the values of \( \rho_1 \) and \( \rho_2 \) are changed such that \( \rho_1 = 0.5, \rho_2 = 0.5 \), then as shown from Table 5.2, the percentage deviation of \( \langle n >'_2, W'_2 \) from \( \langle n >_2, W_2 \) increases. For \( \langle n >'_2 \) the maximum percentage deviation becomes 5.6 while for \( W'_2 \) it becomes 15.2. This indicates that the deviation increases when \( \rho_2 \) decreases.

(iii) If \( \rho_1, \rho_2 \) are further decreased then Table 5.3 shows that the maximum percentage deviation of \( \langle n >'_2 \) from \( \langle n >_2 \) is 9.41 while it is 71.53 for \( W'_2 \). The deviation decreases rapidly as \( C_{s2}^2 \) increases.
The above examples show that for high values of $\rho$, replacement of the $G/GE/1$ by the $GE/GE/1$ has no serious effects on the performance metrics. The percentage difference is particularly negligible when high values of $\rho$ are coupled with high values of the coefficients of variation. For low values of $\rho_2$ the percentage difference is more considerable, especially when $\frac{\sigma^2}{\mu_i}$ are less than one. However, the results obtained using the $G/GE/1$ are still tolerable if $\frac{\sigma^2}{\mu_i}$ are high. In general, the mean waiting time is more affected than the average number of customers.

### 5.3.2 A general two-station cyclic queueing model

Fig. 5.4 shows a two-station cyclic queueing model. In this model there is a finite number of waiting positions $N-1$ and consequently a finite number of customers $N$ are allowed to cycle around the model taking service from station 1 and station 2 respectively as they move. When a customer leaves the system it is immediately replaced by a new customer so that the number of customers in the model remains constant. Assume that service time distributions at both stations are of the GE-type with mean rates $\mu_i$ and coefficients of variation $\frac{\sigma^2}{\mu_i}$ not necessarily equal to one. A typical example is a multiprogramming computer system where $N$ represents the degree of multiprogramming, station 1 represents the CPU and station 2 represents the IO unit. If station 1 is busy, then the interdeparture times for this station are identical to its service times, otherwise all the $N-1$ customers are queued at station 2 and no
Table 5.4a The utilisation of station 2 in the cyclic model of example 5.3.2(i)

<table>
<thead>
<tr>
<th>$1/\mu_2$</th>
<th>0.25</th>
<th>0.5</th>
<th>0.75</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>S</td>
<td>D</td>
<td>M</td>
</tr>
<tr>
<td>2</td>
<td>0.233</td>
<td>0.225</td>
<td>0.226</td>
</tr>
<tr>
<td>3</td>
<td>0.240</td>
<td>0.237</td>
<td>0.238</td>
</tr>
<tr>
<td>4</td>
<td>0.243</td>
<td>0.244</td>
<td>0.244</td>
</tr>
<tr>
<td>5</td>
<td>0.248</td>
<td>0.247</td>
<td>0.247</td>
</tr>
<tr>
<td>6</td>
<td>0.249</td>
<td>0.248</td>
<td>0.249</td>
</tr>
<tr>
<td>7</td>
<td>0.248</td>
<td>0.249</td>
<td>0.249</td>
</tr>
<tr>
<td>8</td>
<td>0.249</td>
<td>0.250</td>
<td>0.250</td>
</tr>
<tr>
<td>9</td>
<td>0.245</td>
<td>0.250</td>
<td>0.250</td>
</tr>
<tr>
<td>10</td>
<td>0.255</td>
<td>0.250</td>
<td>0.250</td>
</tr>
</tbody>
</table>

$S = $ Simulation; $D = $ diffusion approximation; $M = $ maximum energy using a GE-type service time distribution

$C_{S2}^2 = 5; \quad C_{a2}^2 = 1$
Table 5.4b

The average number of customers in station 2 of the cyclic model of example 5.3.2(i)

$c_{a2}^2 = 1, \quad c_{s2}^2 = 5$

<table>
<thead>
<tr>
<th>$1/\mu_2$</th>
<th>$\nu$ = 0.25</th>
<th>$\nu$ = 0.5</th>
<th>$\nu$ = 0.75</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.323</td>
<td>0.615</td>
<td>0.835</td>
</tr>
<tr>
<td>3</td>
<td>0.397</td>
<td>0.855</td>
<td>1.223</td>
</tr>
<tr>
<td>4</td>
<td>0.441</td>
<td>1.057</td>
<td>1.592</td>
</tr>
<tr>
<td>5</td>
<td>0.467</td>
<td>1.227</td>
<td>1.943</td>
</tr>
<tr>
<td>6</td>
<td>0.481</td>
<td>1.369</td>
<td>2.278</td>
</tr>
<tr>
<td>7</td>
<td>0.490</td>
<td>1.488</td>
<td>2.596</td>
</tr>
<tr>
<td>8</td>
<td>0.494</td>
<td>1.586</td>
<td>2.899</td>
</tr>
<tr>
<td>9</td>
<td>0.497</td>
<td>1.666</td>
<td>3.187</td>
</tr>
<tr>
<td>10</td>
<td>0.498</td>
<td>1.732</td>
<td>3.460</td>
</tr>
</tbody>
</table>
Table 5.5a The utilisation of station 2 in the cyclic model of example 5.3.2(ii)

<table>
<thead>
<tr>
<th>N</th>
<th>1/\mu_2</th>
<th>0.25</th>
<th>0.5</th>
<th>0.75</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S</td>
<td>D</td>
<td>M</td>
<td>S</td>
</tr>
<tr>
<td>2</td>
<td>0.221</td>
<td>0.236</td>
<td>0.238</td>
<td>0.412</td>
</tr>
<tr>
<td>3</td>
<td>0.240</td>
<td>0.246</td>
<td>0.247</td>
<td>0.443</td>
</tr>
<tr>
<td>4</td>
<td>0.252</td>
<td>0.249</td>
<td>0.249</td>
<td>0.454</td>
</tr>
<tr>
<td>5</td>
<td>0.242</td>
<td>0.250</td>
<td>0.250</td>
<td>0.475</td>
</tr>
<tr>
<td>6</td>
<td>0.245</td>
<td>0.250</td>
<td>0.250</td>
<td>0.477</td>
</tr>
<tr>
<td>7</td>
<td>0.250</td>
<td>0.250</td>
<td>0.250</td>
<td>0.488</td>
</tr>
<tr>
<td>8</td>
<td>0.247</td>
<td>0.250</td>
<td>0.250</td>
<td>0.499</td>
</tr>
<tr>
<td>9</td>
<td>0.250</td>
<td>0.250</td>
<td>0.250</td>
<td>0.495</td>
</tr>
<tr>
<td>10</td>
<td>0.244</td>
<td>0.250</td>
<td>0.250</td>
<td>0.490</td>
</tr>
</tbody>
</table>

S = simulation;  D = diffusion;  M = maximum entropy using a GE-type service and interarrival time distribution

\[ c_{s2}^2 = 5, \quad c_{a2}^2 = c_{s1}^2 = 0 \]
Table 5.5b The average number of customers in station 2 of the cyclic model of example 5.3.2(ii)

\[
\begin{align*}
C_{a2} & = 0.0 \\
C_{s2} & = 5
\end{align*}
\]

<table>
<thead>
<tr>
<th>1/μ₂</th>
<th>N</th>
<th>0.25</th>
<th>0.5</th>
<th>0.75</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.286</td>
<td>0.600</td>
<td>0.831</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.318</td>
<td>0.813</td>
<td>1.209</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.328</td>
<td>0.980</td>
<td>1.566</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.332</td>
<td>1.110</td>
<td>1.901</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.333</td>
<td>1.211</td>
<td>2.215</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.333</td>
<td>1.287</td>
<td>2.510</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.333</td>
<td>1.344</td>
<td>2.787</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>0.333</td>
<td>1.387</td>
<td>3.045</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.333</td>
<td>1.418</td>
<td>3.286</td>
<td></td>
</tr>
</tbody>
</table>
extra customers are allowed. In this sense station 2 behaves as if it were a GE/GE/1/N queueing system and can be analysed using the results of section 5.2 above with a traffic intensity $\rho_2 = \mu_1/\mu_2$. The effective utilisation of this station is then given by

$$\rho_2' = 1 - p_N(0)$$

while that of station 1 is given by [ALLE 78],

$$\rho_1' = 1 - p_N(N)$$

The average number of customers at station 2 can be obtained using (5.8) and that of station 1 can be obtained using the relation

$$<n>_1 = N - <n>_2$$

Other performance metrics can be obtained using Little's formula in a manner similar to [ALLE 78].

To illustrate the credibility of this approach two numerical examples are considered.

(i) In the first case $\rho_2 = \mu_1/\mu_2$ is varied between 0.25 - 0.75 at steps of 0.25. $N$ is varied between 2 and 10; $C_{s1}^2 = 1$, $C_{s2}^2 = 5$. In each case the effective utilisation $\rho_2'$ is computed and compared with simulation and diffusion results [BADE 79]. The results are displayed in Table 5.4a. It is surprising that the maximum entropy results are identical to those obtained using the diffusion approximation.
Table 5.6

The utilisation and the average number of customers in the $M/H_2/1/N$ for different values of $k$ in the expression for $a_1$ (3.41b).

$N = 4, \ \lambda/\mu = 0.75$

<table>
<thead>
<tr>
<th>$k$</th>
<th>utilisation</th>
<th>$&lt;n&gt;$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.606</td>
<td>1.483</td>
</tr>
<tr>
<td>10</td>
<td>0.581</td>
<td>1.562</td>
</tr>
<tr>
<td>20</td>
<td>0.5788</td>
<td>1.576</td>
</tr>
<tr>
<td>GE-type($k=\infty$)</td>
<td>0.576</td>
<td>1.592</td>
</tr>
<tr>
<td>simulation($k=2$)</td>
<td>0.607</td>
<td></td>
</tr>
</tbody>
</table>
In Table 5.4b the average number of customers at station 2 is listed for the different values of N and \( \rho \).

(ii) The same procedure is applied to the case where \( C_{s1} = 0.0 \) while all the other parameters remain as in the above example. The results are tabulated in Table 5.5a and 5.5b. Again, the maximum entropy results are very close to the diffusion approximation results.

Note that the simulation results were obtained using a hyper-exponential service time [BADE 79]:

\[
f(t) = \alpha_1 \mu_1 e^{-\mu_1 t} - (1-\alpha_1) \mu_2 e^{-\mu_2 t}
\]

while from Chapter 3, \( \alpha_1 \) is given by

\[
\alpha_1 = \frac{C_s^2 - 1}{2(C_s^2 + 1)} + \frac{2}{k(C_s^2 + 1)} + \frac{(C_s^2 - 1)^2 + 8(C_s^2 - 1)/k + 8(1-C_s^2)/k^2}{2(C_s^2 + 1)}
\]

where \( k \) is an arbitrary parameter. For the given simulation results, \( k \) was arbitrarily taken to be 2. If the value of \( k \) is increased, then the simulation results are expected to approach the maximum entropy results. When \( k \to \infty \) the results should be identical. This fact is illustrated in Table 5.6 for some of the results obtained in Table 5.5a.

In Table 5.6 exact results for the M/H_2/1 are computed for different values of \( k \). Therefore, the discrepancy between the simulation and maximum entropy results may be mainly due to the arbitrary parameter \( k \).
5.4 Correlation in the Departure Process

The above analysis has been carried out assuming that the inter-departure times from a G/G/1 queueing system are independent and identically distributed random variables. However, it has been shown [JENK 66] that this assumption may only be true in the case of exponential servers and Poisson arrivals. In all the other cases interdeparture times may exhibit a degree of serial correlation.

Jenkins [JENK 66] studied the particular case of the M/E_k/1 (Poisson arrivals, Erlang service times). For this system he computed the autocorrelation of lag 1 and lag 2 and showed that this autocorrelation is always positive for 0<\rho<1. The effect of this autocorrelation on the performance metrics has been experimentally investigated by Shimshak and Sphicas [SHIM 82]. They observed that this correlation leads to performance metrics lower than those obtained from simulation. However, the problem still requires more consideration.

The results obtained by Jenkins can easily be generalised to study the correlation in the departure process from the maximum entropy M/G/1. The basic result obtained by Jenkins is that the joint density function for departure intervals d_n, d_{n-1} is given by

\[ h_{d_{n-1},d_n}(x,y) = \Pi_o b(x)f(y) + \Pi_o \{b(y)-f(y)\}\lambda e^{-\lambda x} F(x) \]

\[ + \Pi_1 \{b(y)-f(y)\} e^{-\lambda x} f(x)+(1-\Pi_o)f(x)f(y) \]  \hspace{1cm} (5.23)

where f(t) is the service time distribution, \( \Pi_n \) is the distribution of the number of customers in the system imbedded at arrival instants and b(t) is the convolution
Table 5.7 Correlation in the Departure Process

<table>
<thead>
<tr>
<th>$c_a^2$</th>
<th>$c_s^2$</th>
<th>$\rho$</th>
<th>$c_d^2$</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>0.1</td>
<td>4.6</td>
<td>0.0085</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>4.2</td>
<td>0.0316</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.3</td>
<td>3.8</td>
<td>0.0639</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.4</td>
<td>3.4</td>
<td>0.1000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>3.0</td>
<td>0.1356</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.6</td>
<td>2.6</td>
<td>0.1659</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.7</td>
<td>2.2</td>
<td>0.1850</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.8</td>
<td>1.8</td>
<td>0.1828</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.9</td>
<td>1.4</td>
<td>0.1387</td>
</tr>
</tbody>
</table>
\[ b(t) = \int_0^t \lambda e^{-\lambda(t-x)} f(x) \, dx, \quad t \geq 0 \] (5.24)

(x and y are not to be confused with our previous definitions).

It can be seen that (5.13) actually applies to any M/G/1.

For the maximum entropy M/G/1, the autocorrelation of lag 1 can be obtained using the relation:

\[
\text{Corr}(d_{n-1}, d_n) = \frac{E(d_{n-1}, d_n) - E(d_n)E(d_{n-1})}{\sigma_{d_n} \sigma_{d_{n-1}}} \tag{5.25}
\]

but \( E(d_n) = E(d_{n-1}) = \frac{1}{\lambda} \)

Similarly,

\[
\sigma_{d_n} \sigma_{d_{n-1}} = \frac{c^2_d \lambda}{2}
\]

and

\[
E(d_{n-1}, d_n) = \int \int xy h_{d_{n-1}, d_n}(x, y) \, dx \, dy \quad \text{for} \quad 0 < n < \infty
\]

Using (3.15) for \( f(t) \), this gives

\[
E(d_{n-1}, d_n) = \frac{1}{\lambda^2} - \frac{\Pi_0 r_s}{(\lambda + r_s \mu)^2} + \frac{\Pi_1}{\lambda} \left( \frac{1-r_s}{\lambda} + \frac{r_s^2 \mu}{(\lambda + r_s \mu)^2} \right)
\]

since

\[
\Pi_0 = p_0 = 1 - \rho \quad \text{and} \quad \Pi_1 = p_1 = \frac{(1-\rho)r_s(1+y_s)}{1+\rho y_s} \quad \text{(using (3.9))},
\]

the above equation becomes

\[
E(d_{n-1}, d_n) = \frac{1}{\lambda^2} + \frac{(1-r_s)(1-\rho)}{\lambda^2 (1+\rho y_s)} + \frac{r_s(1-\rho)}{(\lambda + r_s \mu)^2} \frac{r_s}{1+\rho y_s} - 1 \tag{5.26}
\]
In view of this, (5.25) becomes

\[
\text{Corr}(d_{n-1}, d_n) = \frac{(1-r_s)(1-\rho)\lambda + \frac{r_s(1-\rho)}{(\lambda+r_s\mu)^2} \frac{r_s}{1+\rho y_s} - 1}{C_d^2} \lambda^2
\]  

(5.27)

where \(C_d^2\) is obtainable from theorem (4.6).

In Table 5.7 correlation in the departure process for an example M/G/I is given.

5.5 Summary

In this chapter, the results previously obtained for the G/GE/I have been utilised to analyse two-server tandem and cyclic queues.

The main purpose of analysing tandem queues has been to evaluate the effect of approximating the interdeparture time distribution by a GE-type distribution. Three examples of infinite capacity tandem queues covering wide ranges of \(\rho\), \(C_a^2\), \(C_s^2\) have been considered. The numerical results displayed in Tables 5.1-5.3 showed that this approximation is reasonable for medium and high values of \(\rho\). The error is expected to be less significant for finite capacity queueing systems.

The analysis of two-server closed cyclic models required the determination of the maximum entropy solution for the G/GE/I/N system. This has been done by making the reasonable assumption that this solution is proportional to the maximum entropy solution of the infinite capacity G/GE/I. The obtained solution is shown in Box. 5.1. The G/GE/I/N is then used to analyse two examples of closed cyclic models.
The numerical results displayed in Tables 5.4-5.5 show that the system utilisation obtained this way is very close to that obtained using the diffusion approximation and simulation methods. This stresses further the similarity between the diffusion and maximum entropy methods. In Table 5.6 it has been illustrated that the simulation results are expected to be identical to the maximum entropy results if the parameters of the hyperexponential used are changed \((k \rightarrow \infty)\).

Finally, the correlation in the departure process has been discussed and evaluated for an example \(M/GE/1\) system.
CHAPTER VI

MAXIMUM ENTROPY ANALYSIS OF GENERAL QUEUEING NETWORKS

6.1 Introduction

The last chapter has been devoted to queueing networks in which job routing is of a restricted nature. In this chapter, networks of arbitrary Markovian routing are considered. Jobs leaving a service centre $i$ in the network can join any other service centre $j$ with probability $q_{ij}$, $0 \leq i, j \leq M$, where $M$ is the total number of service centres, $0$ represents the outside environment and $\sum_{j=0}^{M} q_{ij} = 1$. The evolution of these networks has been summarised in Chapter I.

In the following study a basic maximum entropy model is introduced for general queueing networks. It is then discussed how this model can be implemented in conjunction with Markovian and operational analysis. Since the implementation may often require the determination of the interinput and interdeparture processes, new simplified methods for the approximation of flow processes are described. The credibility of the analysis is illustrated by some applications and examples. The analysis is generally aimed at solving the problems mentioned in Chapter I.

The maximum entropy model is given in section 6.2. Implementation of the model is discussed in section 6.3. Applications and examples are given in section 6.4.
6.2 Maximum Entropy and General Queueing Networks

In this section a basic maximum entropy model is introduced for general queueing networks. The implementation of this model will be considered in the subsequent sections. First, the basic model is described.

6.2.1 Description of the Basic Model

Fig. 6.1 shows a queueing network that consists of $M$ service centres and satisfies the following conditions:

(i) Routing in the network is described by a transition matrix $Q = \{q_{ij}\}$, $0 < i, j < M$ where $q_{oi}$ is the probability that an external arrival is directed to server $i$, $q_{io}$ is the probability that a job finishing service at centre $i$ leaves the system and $q_{ij}$, $1 < i, j < M$, is the probability that a job joins centre $j$ having just completed service at centre $i$. Further assume that each device is reachable from any other device.

(ii) For each device $i$, $1 < i < M$ the mean service rate is $\mu_i$ and the coefficient of variation for the service time is $C_{\mu_i}^2$.

(iii) For external arrivals the mean rate is $\lambda_o$ and the coefficient of variation is $C_{\lambda_o}^2$.

Note that at this stage we are not committing ourselves to any assumptions about the service or interarrival time distribution.

Defining $A_i$ to be the expected number of visits per job to centre $i$, then $A_i$, $1 < i < M$, satisfies

$$A_i = q_{oi} + \sum_{j=1}^{M} A_j q_{ji}$$

(6.1)

Based on this definition, the mean arrival rate to device $i$ will be given by
Fig. 6.1 An example of a general queueing network
\[ \lambda_i = \lambda_o \Lambda_i \]  

(6.2)

The utilisation of server \( i \), \( \rho_i \) is given by

\[ \rho_i = \frac{\lambda_o \Lambda_i}{\mu_i} \text{ if } \lambda_o \Lambda_i < \mu_i \]  

(6.3)

The state of the network can be described by the integer valued vector \( \mathbf{n} = (n_1, n_2, \ldots, n_M) \), \( n_i > 0, 1 \leq i \leq M \). Let \( p(n) \) be the equilibrium probability that the system is in state \( \mathbf{n} \). The marginal equilibrium probability that centre \( i \) is in state \( n_i \) is given by

\[ p_i(n_i) = \sum_{\text{all } n \text{ with } n_i \text{ fixed}} p(n) \]  

(6.4)

This type of model is known in the literature as open queueing networks.

If in this model the number of customers is fixed and both \( q_{oi} \) and \( q_{io} \) are set to zero then we have a closed queueing network for which (6.1) becomes

\[ \Lambda_i = \sum_{j=1}^{M} \Lambda_j q_{ji} \]  

(6.5)

(6.5) has no unique solution and (6.3) can no longer apply. (6.5) is usually solved in this case by fixing one of the \( \Lambda_i \).

6.2.2 A maximum entropy model

Open queueing networks

Theorem 6.1

For the open queueing network described above, if the average number of customers \( \langle n \rangle_i \) at each centre is finite, then the maximum
entropy probability distribution that the network is in state \( n = (n_1, n_2, \ldots, n_M) \) is given by

\[
p(n) = \prod_{i=1}^{M} p_i(n_i) \tag{6.6}
\]

where \( p_i(n_i) \) is the marginal maximum entropy probability distribution that service centre \( i \) is in state \( n_i \), and is given by

\[
p_i(n_i) = \begin{cases} 
(1 - \rho_i) & n_i = 0 \\
(1 - \rho_i)g_i x_i^n & n_i > 0
\end{cases} \tag{6.7}
\]

where

\[
x_i = \frac{\langle n_i \rangle - \rho_i}{\langle n \rangle} \quad \text{and} \quad g_i = \frac{\rho_i (1 - x_i)}{(1 - \rho_i) x_i} \tag{6.8}
\]

Proof

For the above model, the maximum entropy solution can be obtained by maximising (2.6) written as:

\[
H(p) = - \sum_{\text{all } n} p(n) \ln p(n)
\]

subject to the constraints

\[
\sum_{\text{all } n} p(n) = 1
\]

\[
\sum_{n_i=0}^{\infty} L_i(n_i)p_i(n_i) = \rho_i, \quad \text{with } L_i(n_i) = \begin{cases} 0 & n_i = 0 \\
1 & n_i > 0
\end{cases} \tag{6.9}
\]

and

\[
\sum_{n_i=0}^{\infty} n_i p_i(n_i) = \langle n_i \rangle, \quad 1 \leq i \leq M \tag{6.10}
\]

which gives by (2.13)

\[
\text{Note that } \sum_{n_i=0}^{\infty} L_i(n_i)p_i(n_i) = \sum_{\text{all } n} L_i(n_i)p(n) \quad \text{and} \quad \sum_{n_i=0}^{\infty} n_i p_i(n_i) = \sum_{\text{all } n} n_i p(n)
\]
\[
p(n) = Z_p^{-1} \exp\left\{ -\sum_{i=1}^{M} (\beta_{i1} L_1(n_i) + \beta_{i2} n_i) \right\}
\] (6.11)

and by (2.14) \(Z_p\) is given by

\[
Z_p = \sum_{\text{all } n} \exp \left\{ -\sum_{i=1}^{M} (\beta_{i1} L_1(n_i) + \beta_{i2} n_i) \right\}
\]

\(\beta_{i1}\) and \(\beta_{i2}\) are the Lagrange multipliers corresponding to the constraints (6.9) and (6.10).

Defining \(g_i = \exp(-\beta_{i1})\), and \(x_i = \exp(-\beta_{i2})\), \(Z_p\) can be expressed as

\[
Z_p = \sum_{\text{all } n} \prod_{i=1}^{M} g_i^{L_i(n_i)} x_i^{n_i}
\]

Interchanging the sum and product in this expression, one gets

\[
Z_p = \prod_{i=1}^{M} \left( \sum_{n_i} g_i^{L_i(n_i)} x_i^{n_i} \right)
\]

The bracketed term is nothing but the expression for \(Z_{p_i}\) obtained if the maximum entropy solution \(p_i(n_i)\) is derived separately for each node \(i\). Therefore

\[
Z_p = \prod_{i=1}^{M} Z_{p_i}
\]

and consequently (6.11) can be expressed as

\[
p(n) = \prod_{i=1}^{M} Z_{p_i}^{-1} g_i^{L_i(n_i)} x_i^{n_i} = \prod_{i=1}^{M} Z_{p_i}(n_i)
\]

Considering each node separately the values of \(x_i\) and \(g_i\) given in (6.8) can be obtained (see Chapter 4).

Q.E.D.
At this stage the maximum entropy model of theorem (6.1) is independent of any stochastic process and is defined only on the actual values of $p_i$ and $<n>_i$.

Closed Queueing Networks

For $f_{\pi_{\infty}}$ queueing networks, the above arguments lead to a maximum entropy solution similar in form to (6.11). Defining $x_i$ and $g_i$ as above, this solution can be expressed as

$$p_N(n) = Z_p^{-1} \prod_{i=1}^{M} g_i x_i$$

where $N$ is the total number of customers in the network (which is fixed in this case). However, in this case $Z_p$ cannot be decomposed into the $Z_{pi}$ because of the constraint

$$\sum_{i=1}^{M} n_i = N$$

Based on this, and the results obtained in Chapter 5 for closed cyclic models, one can conjecture that for closed queueing networks

$$p_N(n) = C_p(n)$$

where $C$ is a proportionality constant. Using (6.6) and (6.7) this becomes

$$p_N(n) = C^{-1} \prod_{i=1}^{M} g_i x_i$$

where $C$ is a normalising constant, given by

$$C = \prod_{\text{all } n} \left( \frac{N}{\sum n_i} \right)^{\frac{1}{E}}$$

and

$$L_i(n_i) = \begin{cases} 0 & n_i = 0 \\ 1 & n_i > 1 \end{cases}$$
For the open network described in section 6.2.1 the maximum entropy solution is given by

\[
p(n) = \prod_{i=1}^{M} p_i(n_i)
\]

where

\[
p_i(n_i) = \begin{cases} \frac{(1-p_i)}{n_i} & n_i = 0 \\ \frac{n_i}{(1-p_i)} g_i x_i & n_i > 0 \end{cases}
\]

For the closed network, the maximum entropy solution is given by

\[
p(n) = G^{-1} \prod_{i=1}^{M} g_i x_i
\]

where \( G \) is a normalising constant given by:

\[
G = \sum_{\text{all } n_i \geq 0} \prod_{i=1}^{M} L_i(n_i) \frac{n_i}{x_i}
\]

with \( 0 < n_i < N \), and \( \text{sum}(n_i) = N \)

\[
L_i(n_i) = \begin{cases} 0 & n_i = 0 \\ 1 & n_i > 0 \end{cases}
\]
6.3 Implementation of the Maximum Entropy Solution

6.3.1 Maximum entropy and Markovian queueing analysis

For the maximum entropy model of theorem 6.1, if the actual value of \( \rho_i \) and \( \langle n \rangle_i \) are provided, then the solution can be completely specified. However, in practice, the analyst is faced with the reverse problem of predicting the value of \( \langle n \rangle_i \) and the other performance metrics depending on it. In such cases it is often useful to approximate the actual behaviour of the system by a suitable stochastic process. The stochastic process can then be a good approximation for the actual maximum entropy solution.

One such approximation is to assume that for each service centre \( i \), service times are independent and identically distributed random variables with mean \( \mu_i \), coefficient of variation \( C_{si} \) and a GE-type pdf written as

\[
f_i(t) = (1-r_{si})u_0(t) + r_{si}u_1 \exp \{ -\mu_1 t \}
\]

where \( r_{si} = 1/(1+\gamma_{si}) \), \( \gamma_{si} = (C_{si}^2-1)/2 \) and \( \mu_1' = r_{si} \mu_1 \).

If arrivals at each node are also assumed to be independent and identically distributed random variables, (ignoring any correlation in the interarrival process) then the average number of customers at each node \( i \) in the open network will be given by (4.23) as:

\[
\langle n \rangle_i = \frac{\lambda_i}{\theta_{1i} - \rho_i \gamma_{si}}
\]

where \( \lambda_i \) and \( \rho_i \) are given by (6.2) and (6.3) respectively. This follows since under these assumptions each node will behave as a \( G/GE/1 \) system. As in Chapter 4, \( \theta_{1i} \) is the zero of the polynomial
\[ \dot{\mu}_i + \theta - (\mu_i' + (1-r_{si}) \theta) A_i^* (-\theta) = 0 \]

where \( A_i^* (-\theta) \) is the Laplace transform of the interinput distribution for node \( i \).

Substituting (6.14) in (6.6)-(6.8) we get

\[ p(n) = \sum_{i=1}^{M} (1-\rho_i) \frac{\mu_i - \theta (1+y_{si}) n_i}{\mu_i - y_{si} \theta_{li}} n_i \quad n \neq 0 \]

where

\[ g_i = \frac{\rho_i (1-x_i)}{(1-\rho_i) x_i} \quad (6.16) \]

For closed network this solution becomes

\[ p_{N}(n) = \sum_{i=1}^{M} \frac{\mu_i - \theta_{li} (1+y_{si}) n_i}{\mu_i - y_{si} \theta_{li}} n_i \]

where \( G \) is a normalising constant chosen such that

\[ \sum p(n) = 1 \]

and \( \text{sum}(n_i) = N \).

\( g_i \) is given by (6.16).

There are two problems with the solutions (6.15)-(6.17).

a. Because of the complexity of the flow process it may be very difficult to obtain the actual value of \( \theta_{li} \).

b. Even if the flow processes can be determined, it is known that this flow is correlated \(|\text{DISN 77}|,|\text{GELE 80}|\). This correlation is zero only for Poisson flow (see section 5.4).
Box 6.2

If we assume a GE-type distribution for each centre, then the maximum entropy solution for the open network becomes:

$$p(n) = \left\{ \begin{array}{ll}
\prod_{i=1}^{M} (1-\rho_i) & n=0 \\
\frac{\prod_{i=1}^{M} \left( \frac{\mu_i - \theta_i (1+y_{s_i})}{\mu_i - y_{s_i} \theta_i} \right)^n_i}{\prod_{i=1}^{M} (1-\rho_i) g_i \left( \frac{\mu_i - \theta_i (1+y_{s_i})}{\mu_i - y_{s_i} \theta_i} \right)^n_i} & n \neq 0
\end{array} \right.$$

For the closed network:

$$p_n(n) = \frac{\prod_{i=1}^{M} \left( \frac{\mu_i - \theta_i (1+y_{s_i})}{\mu_i - y_{s_i} \theta_i} \right)^n_i}{\gamma - 1 \prod_{i=1}^{M} g_i \left( \frac{\mu_i - \theta_i (1+y_{s_i})}{\mu_i - y_{s_i} \theta_i} \right)^n_i}$$

where $\gamma$ is a normalising constant that can be obtained using the Buzen algorithm.
However, despite these problems, it is shown in the subsequent sections that this solution gives satisfactory results.

Note that this solution implies that the individual nodes will behave as if they were independent M/M/1 systems only if the flow is purely Poisson, which is the case in feedforward Jacksonian networks. In feedback Jacksonian networks, the individual nodes will behave as if they were independent G/M/1 systems rather than M/M/1 systems.

6.3.2 Flow in the network

If the solutions (6.15) and (6.17) are to be implemented, then the flow processes in the network should be characterised. These are mainly the interdeparture and the interinput time distributions. It has already been discussed (section 5.3) how these distributions can be determined in the case of simple tandem queues. However, for networks with general Markovian routing, these distributions are more difficult to obtain and pose a serious problem in the computation of the performance metrics. This problem has previously been recognised by Reiser and Kobyashi [REIS 75], Gelenbe and Pujolle [GELE 76], Sevick et al [SEVI 77] and Pujolle and Soula [PUJO 81]. Their discussions have been in the context of the diffusion approximation or the aggregation of general queueing networks.

In this section an algorithm that computes the squared coefficients of variation for the interinput and the interdeparture distributions is derived. To this end we make the simplifying assumption that the interdeparture time pdf at each node i is given by
\[ d_i(t) = (1 - r_{di})u_0(t) + r_{di} \lambda_i e^{-\lambda_i t} \]  
\[ (6.18) \]

and the PDF is

\[ D_i(t) = 1 - r_{di} e^{-\lambda_i t} \]  
\[ (6.19) \]

where

\[ r_{di} = \frac{2}{(C_{di}^2 + 1)} \quad \text{and} \quad \lambda_i' = r_{di} \lambda_i \]  
\[ (6.20) \]

and \( C_{di}^2 \) is the squared coefficient of variation of the interdeparture time distribution. As has been shown in section 5.3.1, this assumption is reasonable for medium or high values of \( \rho_i \) and especially good when \( C_{si}^2 > 1 \).

The interinput distribution for each node \( i \) can then be obtained by merging the interdeparture distributions from all the sources \( j \) such that \( p_{ji} > 0 \). The merged stream can be obtained by successively applying an algorithm for two-way merging introduced in [SEVI 77], p.11, which implies that:

Given any two streams of random variable \( X_1, X_2 \), each of distribution function

\[ B_i(t) = P(X_i \leq t) \quad i = 1, 2 \]

then the distribution function of the joint stream \( X \), say, is given by

\[ V_2(t) = P(X \leq t) = \frac{E(X_2)H_1(t) + E(X_1)H_2(t)}{E(X_1)E(X_2)} \]  
\[ (6.21) \]

where \( H_1(t) \) is given by

\[ H_1(t) = B_1(t) + \frac{(1-B_1(t))}{E(X_2)} \int_0^t (1-B_2(u)) \, du \]  
\[ (6.22) \]
and similarly

$$H_2(t) = B_2(t) + \int_0^t (1 - B_1(\nu)) d\nu$$  \hspace{1cm} (6.23)

Applying these equations to the queueing network and setting $B_i(t) = D_i(t)$, $i = 1, 2$, we get (see Fig. 6.2):

$$v_2(t) = 1 - \frac{(\lambda_1' + \lambda_2')}{\lambda_1 + \lambda_2} \exp \{- (\lambda_1' + \lambda_2') t\}$$

Applying this procedure successively to all queues $j$ such that $p_{ji} > 0$, we get:

$$A_i(t) = v_M(t) = 1 - \left( \sum_{k=1}^{M_i} \frac{\lambda_k'}{\lambda_k} \right) \exp \{- \sum_{k=1}^{M_i} \lambda_k' t\} \hspace{1cm} (6.24)$$

where $M_i$ is the number of nodes such that $p_{ji} > 0$.

Note that $\lambda_i' = r_i \lambda_i$ and $\lambda_i$ is given by 6.2

Defining

$$\gamma_{1i} = \sum_{k=1}^{M} P_{ki} \lambda_k' \hspace{1cm} \text{and} \hspace{1cm} \gamma_{2i} = \sum_{k=1}^{M} P_{ki} \lambda_k$$

we get

$$A_i'(t) = 1 - \frac{\gamma_{1i}}{\gamma_{2i}} \exp \{- \gamma_{1i} t\} \hspace{1cm} (6.26)$$

which is the interinput PDF for queue $i$. From this the pdf of the interinput process is obtained as
Fig. 6.2 Merging two streams
Therefore the first moment of the interinput pdf is

\[ E(X) = \frac{1}{\gamma_{2i}} \]  \hspace{1cm} (6.28)

and the second moment is

\[ E(X^2) = \frac{2}{\gamma_{1i} \cdot \gamma_{2i}} \]  \hspace{1cm} (6.29)

from which the squared coefficient of variation is

\[ C_{a_i}^2 = \frac{E(X^2)}{E(X)} - 1 \]

\[ = \frac{(2\gamma_{2i}/\gamma_{1i})}{\gamma_{2i}} - 1 \]

\[ = 2(\frac{\sum_{k=1}^{M_i} \rho_{ki} \lambda_k}{\sum_{k=1}^{M_i} \rho_{ki} \lambda_k}) - 1 \]  \hspace{1cm} (6.30)

where \( \lambda_k' = r_{d_k} \lambda_k \), \( r_{d_k} = 2/(C_{a_k}^2 + 1) \)

and \( C_{d_k}^2 \) is the squared coefficient of variation of the interdeparture process for queue \( k \), which can be obtained using theorem (4.6) and theorem (4.7) as:

\[ C_{d_k}^2 = \rho_k^2(C_{s_k}^2 + 1) + (1-\rho_k)(2\rho_k + 1 + C_{a_k}^2) - 1 \]  \hspace{1cm} (6.31)

Clearly \( C_{d_k}^2 \) depends also on the interinput distribution coefficients.
of variation $C_{a_k}^2$. More about the derivation of (6.31) can be found in \cite{SEVI77}.

Based on this information the following algorithm can be introduced:

**Algorithm 6.1**

To compute $C_{a_i}^2$ and $C_{d_i}^2$ $1 \leq i \leq M$

1. initialise:

   For $i=1$ to $M$
   
   $C_{d_i}^2 = C_{s_i}^2$

2. repeat

   for $i=1$ to $M$

   BEGIN

   $r_{d_i} = \frac{2}{(C_{d_i}^2 + 1)}$  \quad  $\lambda_i' = r_{d_i} \lambda_i$

   $C_{a_i}^2 = 2 \left( \sum_{k=1}^{M_i} \frac{\lambda_k \cdot k_i}{\lambda_i'} \right) - 1$

   $C_{d_i}^2 = \rho_i \left( C_{s_i}^2 + 1 \right) + (1-\rho_i)(2\rho_i + 1 + C_{a_i}^2) - 1$

   END

   UNTIL $C_{d_i}^2$ converges.

In the following applications, this algorithm will be used to obtain the performance metrics of the individual nodes.
Box 6.3

Algorithm 6.1 To compute $C_{ai}^2$ and $C_{di}^2$ for each node $i$ in the network:

1. initialise:
   
   For $i=1$ to $M$
   
   $C_{di}^2 = C_{si}^2$

2. repeat
   
   for $i=1$ to $M$
   
   BEGIN
   
   $r_{di} = 2/(C_{di}^2 + 1)$ $\lambda_i' = r_{di}\lambda_i$

   $M_i$ $M_i$

   $C_{ai}^2 = 2(\sum_{k=1}^{M} \lambda_k p_{ki}/\sum_{k=1}^{M} \lambda_k' p_{ki}) - 1$

   $C_{di}^2 = \rho_i^2 (C_{si}^2 + 1) + (1-\rho_i)(2\rho_i + 1 + C_{ai}^2) - 1$

   END

   UNTIL $C_{di}^2$ converges.

Note that this algorithm does not take into consideration the split process described by Sevick et al [SEVI 77]. This may be the reason that the results of tables 6.4, 6.5 and 7.1 are inconsistent in comparison with the exact results given by Sevick et al [SEVI 77]. Further investigations are needed.
6.3.3 Maximum entropy and the operational point of view

Another way of implementing the maximum entropy model is to provide direct numerical measures for the required expectations. This is possible if the model is going to be used in consistency checking or performance calculation. There are two options in this case.

The first option is to obtain a measure \( \langle n \rangle_{i_T} \) for the expected number of customers in the system during an observation period \( T \) and obtain the required probabilities accordingly. In this case, there will be no need to compute the flow; the main use of the model here will be to characterise the maximum entropy recursions of section 4.4 and generalise them for any observation period.

The other option is to use equation (4.25) as an estimate for

\[
\lim_{T \to \infty} \langle n \rangle_{i_T} = \frac{1 + \gamma_{ai} + \rho_i \cdot \gamma_{si}}{1 - \rho_i} - \frac{\lambda_i \overline{I}}{2 \overline{I}_i}
\]

and instead of measuring \( \langle n \rangle_{i_T} \) one just measures \( \overline{I} \) and \( \overline{I}^2 \) which are the first and second moments of the idle time distribution. Most of the available monitors provide data about the idle period. However, in this case one needs an estimate for \( C_{ai}^2 \), the squared coefficient of variation for the interinput time distribution.
6.4 Applications and Examples

In this section the credibility of the above theory is illustrated by considering three practical applications.

6.4.1 An open network

Consider the two-server open network shown in Fig. 6.3. In this model it is assumed that the external arrival process is Poisson with mean rate \( \lambda_0 \). The service times at each node are general, independent and identically distributed random variables with mean rates \( \mu_1, \mu_2 \) and coefficients of variation \( C_{s1}, C_{s2} \). This model may be thought of as a computer system with the first node representing the CPU and the second node representing the I/O unit. An arriving job queues for service at node 1 and when this service is completed, the job either leaves the system with probability \( (1-q_1) \) or queues for service at node 2 with probability \( q_1 \). When this service is over, the job either queues again for service at node 2 with probability \( q_2 \) or rejoins the queue at node 1 with probability \( (1-q_2) \). This model has previously been considered by Reiser and Kobyashi [REIS 74], Gelenbe and Pujolle [GELE 76], Pujolle and Soula [PUJO 79], Gelenbe and Mitra [GELE 80]. Here, this model is analysed in three different ways, comparing the results each time with diffusion and simulation:

(i) The interinput distribution to each node is assumed to be of the form

\[
a_{i1}'(t) = (1-r_{a1})u_0(t) + r_{a1}\lambda_1\exp(-\lambda_1't) \quad i=1,2 \tag{6.33}
\]

For station 1, \( C_{a1}^2 \) is estimated using equation (6.30). For station 2 \( C_{a2}^2 \) is estimated from the interinput formula (4.81) derived for the GE/GE/1 with feedback. For both stations \( C_{d1}^2 \) is estimated using formula...
The open network of example 6.4.1
Table 6.1  The average number of customers at each node $i$ of the open network compared with simulation, example 6.4.1(i)

<table>
<thead>
<tr>
<th>Exp. No.</th>
<th>Q. No.</th>
<th>$C_s^2$</th>
<th>$C_a^2$</th>
<th>$C_d^2$</th>
<th>$\rho$</th>
<th>$&lt;n&gt;^M$</th>
<th>% deviation from simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.427</td>
<td>0.958</td>
<td>0.963</td>
<td>0.953</td>
<td>14.05</td>
<td>1.66</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0</td>
<td>0.968</td>
<td>0.978</td>
<td>0.901</td>
<td>4.87</td>
<td>37.8</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.423</td>
<td>0.968</td>
<td>0.982</td>
<td>0.766</td>
<td>2.49</td>
<td>5.55</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0</td>
<td>0.987</td>
<td>0.992</td>
<td>0.713</td>
<td>1.58</td>
<td>15.55</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.414</td>
<td>0.969</td>
<td>0.981</td>
<td>0.646</td>
<td>1.44</td>
<td>10.16</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0</td>
<td>0.986</td>
<td>0.991</td>
<td>0.620</td>
<td>1.11</td>
<td>24.33</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.432</td>
<td>0.980</td>
<td>0.987</td>
<td>0.544</td>
<td>0.998</td>
<td>4.51</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0</td>
<td>0.9907</td>
<td>0.994</td>
<td>0.520</td>
<td>0.796</td>
<td>20.74</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.422</td>
<td>0.971</td>
<td>0.98</td>
<td>0.473</td>
<td>0.761</td>
<td>0.3957</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0</td>
<td>0.985</td>
<td>0.9896</td>
<td>0.446</td>
<td>0.619</td>
<td>15.32</td>
</tr>
</tbody>
</table>

Mean % deviation from simulation  13.60

Mean % deviation of queue 1 = 4.455

Mean % deviation of queue 2 = 22.748
| Parameters | Exp. No. | Q.No. | $\rho_i$ | $q_i$ | $C_{si}^2$ | $C_{ai}^2$ | $C_{di}^2$ | Simulation | Max. Entropy | Diffusion G/P | Diffusion G/P mod. | Diffusion R/K |
|-----------|---------|-------|---------|------|---------|---------|---------|------------|-------------|---------------|----------------|----------------|---------------|
|           |         |       |         |      |         |         |         |            |             |               |                 |                |               |
|           |         | 2     | .901   | .503 | 0       | .650    | .153    | 7.831      | 5.84        | 3.124         | 5.735          | 2.949          |
| 2         |         | 1     | .766   | .509 | .423    | .825    | .621    | 2.359      | 2.240       | 2.106         | 2.093          | 2.002          |
|           |         | 2     | .713   | .499 | 0       | .759    | .422    | 1.871      | 1.800       | 1.031         | 1.694          | 0.970          |
| 3         |         | 1     | .646   | .516 | .414    | .864    | .707    | 1.603      | 1.345       | 1.207         | 1.200          | 1.168          |
|           |         | 2     | .620   | .506 | 0       | .800    | .543    | 1.467      | 1.234       | 0.718         | 1.149          | 0.709          |
| 4         |         | 1     | .544   | .512 | .432    | .899    | .785    | 1.046      | 0.944       | 0.821         | 0.818          | 0.815          |
|           |         | 2     | .520   | .502 | 0       | .857    | .663    | 1.005      | 0.877       | 0.502         | 0.781          | 0.545          |
| 5         |         | 1     | .473   | .504 | .422    | .920    | .824    | 0.758      | 0.736       | 0.619         | 0.617          | 0.632          |
|           |         | 2     | .446   | .507 | 0       | .878    | .740    | 0.731      | 0.668       | 0.382         | 0.589          | 0.453          |

% Mean deviation from simulation 11.28 36.72 18.98 34.59

mean % deviation of queue 1 =

mean % deviation of queue 2 =

G/P = Gelenbe and Pujolle eqn. |GELE 76|

R/K = Reiser and Kobayashi, eqn. |REIS 74|
Table 6.3 The average number of customers at each node $i$ of the open network compared with simulation, example 6.4.1(iii).

<table>
<thead>
<tr>
<th>Exp. No.</th>
<th>Q.No.</th>
<th>$\rho_i$</th>
<th>$C_a^2$</th>
<th>$C_s^2$</th>
<th>$\langle n \rangle$</th>
<th>% Diff. from simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.953</td>
<td>1</td>
<td>0.427</td>
<td>14.49</td>
<td>4.8</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.901</td>
<td>1</td>
<td>0</td>
<td>7.08</td>
<td>9.5</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.766</td>
<td>1</td>
<td>0.423</td>
<td>2.54</td>
<td>7.67</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.712</td>
<td>1</td>
<td>0</td>
<td>2.04</td>
<td>9.09</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.646</td>
<td>1</td>
<td>0.414</td>
<td>1.47</td>
<td>8.29</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.620</td>
<td>1</td>
<td>0</td>
<td>1.37</td>
<td>6.61</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.544</td>
<td>1</td>
<td>0.432</td>
<td>1.01</td>
<td>3.44</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.520</td>
<td>1</td>
<td>0</td>
<td>0.941</td>
<td>6.289</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.473</td>
<td>1</td>
<td>0.422</td>
<td>0.774</td>
<td>2.139</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.446</td>
<td>1</td>
<td>0</td>
<td>0.714</td>
<td>2.32</td>
</tr>
</tbody>
</table>

Mean percentage deviation 6.0135

Mean deviation of queue 1 = 5.267%

Mean deviation of queue 2 = 6.76%
Algorithm 6.1 is then applied until the values of $C_{ai}^2$ and $C_{di}^2$ converge. The results are compared with diffusion and simulation in Table 6.1. (ii) Secondly, the above example is analysed using a modification suggested by Gelenbe and Mitrani [GELE 80]. Namely, if for queue $i$, $q_{ii} > 0$, then $\mu_i$ is replaced by $\mu_i (1-q_{ii})$

$\lambda_i$ is replaced by $\lambda_i (1-q_{ii})$

and $C_{si}^2$ is replaced by $C_{si}^2 (1-q_{ii})+q_{ii}$

and $q_{ij}$ is replaced by

$$q_{ij} = \begin{cases} 0 & \text{if } j=i \\ \frac{q_{ij}}{(1-q_{ii})} & \text{if } j\neq i \end{cases}$$

The purpose of this modification is to include the feedback effect in the analysis indirectly. Algorithm 6.1 is then applied and the average number of customers at each node is computed. The results are shown in Table 6.2 accompanied by % diff. from simulation. (iii) Maintaining the above modification, we assumed that arrivals for both queues are Poisson. The P-K mean value formula is then used to obtain the average number of customers at each node. The results are shown in Table 6.3.

It is interesting to note that the best results are obtained when the flow is assumed to be purely Poisson. The simulation results have been taken from [GELE 80]. No confidence intervals were provided.

6.4.2 A central server model

Consider the central server model shown in Fig. 6A. The parameters of each queue are shown on the Figure. The model is assumed to be closed with a fixed number of customers and FCFS disciplines at all centres.
Fig. 6.4 Example 6.4.2

Fig. 6.5 Example 6.4.2
Table 6.4 Some performance metrics for the central server model of Fig. 6.4

<table>
<thead>
<tr>
<th>Method</th>
<th>Node 1</th>
<th></th>
<th>Node 2</th>
<th></th>
<th>Node 3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>utilis.</td>
<td>&lt;n&gt;</td>
<td>utilis.</td>
<td>&lt;n&gt;</td>
<td>utilis.</td>
<td>&lt;n&gt;</td>
</tr>
<tr>
<td>Exact Sevick et al</td>
<td>0.461</td>
<td>1.068</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>[SEVI 77]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Norton</td>
<td>0.532</td>
<td>1.134</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Approx.1. Sevick et al</td>
<td>0.538</td>
<td>1.115</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Approx.2. Sevick et al</td>
<td>0.538</td>
<td>1.122</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Maximum entropy using Algorithm 6.1</td>
<td>0.468</td>
<td>1.02255</td>
<td>0.588</td>
<td>1.419</td>
<td>0.611</td>
<td>1.558</td>
</tr>
<tr>
<td>Maximum entropy assuming Poisson arrivals to each queue</td>
<td>0.491</td>
<td>0.952</td>
<td>0.623</td>
<td>1.401</td>
<td>0.669</td>
<td>1.644</td>
</tr>
</tbody>
</table>
Table 6.5: Some performance metrics for the central server model of Fig. 6.5

<table>
<thead>
<tr>
<th>Method</th>
<th>Node 1</th>
<th></th>
<th></th>
<th>Node 2</th>
<th></th>
<th></th>
<th>Node 3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>utilis.</td>
<td>&lt;n&gt;</td>
<td></td>
<td>utilis</td>
<td>&lt;n&gt;</td>
<td></td>
<td>utilis</td>
<td>&lt;n&gt;</td>
</tr>
<tr>
<td>Exact Sevick et al</td>
<td>0.573</td>
<td>1.570</td>
<td></td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Norton</td>
<td>0.552</td>
<td>1.581</td>
<td></td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Approx. Sevick et al</td>
<td>0.563</td>
<td>1.585</td>
<td></td>
<td>-</td>
<td>-</td>
<td></td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Maximum entropy using</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Algorithm 6.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maximum entropy using</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Poisson arrivals to each queue</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.622</td>
<td>1.833</td>
<td>0.029</td>
<td>0.040</td>
<td>0.861</td>
<td>3.13</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.686</td>
<td>1.99</td>
<td>0.029</td>
<td>0.030</td>
<td>0.873</td>
<td>2.98</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 6.6  Flow parameters for the models of Figs. 6.4 and 6.5

\(C_a^2\) obtained using algorithm 6.1

<table>
<thead>
<tr>
<th>Model of Fig. 6.4</th>
<th>Model of Fig. 6.5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Node</strong></td>
<td><strong>(C_s^2)</strong></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>
Using (6.17), (6.18) and assuming that the interinput distribution to each node is of the form (6.28), we get (assuming a GE/GE/1):

\[
p(n) = \frac{1}{Z_p^{-1}} \prod_{i=1}^{M} \frac{\rho_i}{\rho_i + y_{ai} + p_i y_{si}} \left( \frac{\rho_i + y_{ai} + p_i y_{si}}{1 + y_{ai} + p_i y_{si}} \right)^{n_i},
\]

where \( \rho_i = \frac{\lambda_i}{\mu_i} \).

Applying Buzen algorithm, the performance metrics shown in Table 6.4. For queue 1 the results are compared with the exact and aggregation results given in Sevick et al. The exact results of Sevick et al were obtained using program QSOLVE. Clearly the maximum entropy results are closest to the exact results obtained by Sevick et al. The procedure is repeated for the model in Fig. 6.5 and the results are shown in Table 6.5.

6.5 Summary

In this chapter a maximum entropy model has been developed for general queueing networks. This model gives the joint maximum entropy probability distribution for the number of customers in an open general network as a product of the marginal maximum entropy solutions for the individual service centres. The maximum entropy solution for the closed network is obtained by making the reasonable assumption that this solution is proportional to that of the corresponding open network (when \( N \to \infty \)). The results are shown in Box 6.1.

It has then been discussed that the maximum entropy model can be implemented either in conjunction with Markovian queueing theory or directly in the 'operational' sense. If the model is to be implemented in the context of Markovian theory, then a GE-type service time
distribution is assumed for each service centre. If the correlation in the departure process is ignored, then each node is expected to behave as if it were a G/GE/1 system. Consequently, the maximum entropy solution shown in Box 6.2 is obtained. However, this solution requires the determination of the flow parameters for the network. An algorithm approximating these parameters is shown in Box 6.3.

If the model is to be applied in an operational sense, there are two options: either $n_i$ is measured directly or, instead, the first and second moments of the idle time distribution are measured in which case the actual value of $n_i$ can be approximated by formula 4.25. In this case one need not consider the correlation unless it is implied by the given data.

To explain the credibility of the analysis, examples of open and closed networks were considered. The results show that the methodology developed compares favourably with diffusion and simulation. The discrepancy between simulation and maximum entropy results may be due to the ignored correlation in the departure process and the fact that the simulation is carried out using a phase-type service time distribution.

It is interesting to note that the coefficient of variation of the interinput time distribution consistently approaches one, reinforcing the findings of Sevick et al |SEVI 77|. This may be explained by the fact that the superimposition of several renewal processes gives rise to a Poisson process |COUR 77|. Note that the best results for the open network are obtained by assuming a Poisson interinput process for each queue.
Another interesting point is that with the maximum entropy solution, the individual nodes will behave as if they were M/M/1 systems if \( C_a^2 = 1 \) and \( C_s^2 = 1 \) which is only the case in feedforward Jacksonian networks. In feedback Jacksonian networks the nodes would behave as if they were independent G/M/1 systems rather than M/M/1 systems, that is if the serial correlation is ignored.
CHAPTER VII
AGGREGATION OF GENERAL QUEUEING NETWORKS

7.1 Introduction

Although the methods described in the previous chapters can now apply to a wide variety of models giving satisfactory results, there are certain circumstances where decomposition of the model is more advantageous e.g. if the state space of the model is very large or if there is simultaneous resource holding in the modelled system, then it is better to decompose the model and analyse each device separately.

In decomposing a queueing network model it is the usual practice to reduce the state space of the network by aggregating the nodes to form composite queues where for each node \( i \) there is a corresponding composite queue. Therefore each node \( i \) can be studied in conjunction with the corresponding composite queue without losing much of the information pertaining to the interaction between this node and the remaining nodes in the network.

There are basically two main approaches for aggregation:

(i) the near-complete decomposability approach introduced by Courtois \([\text{COUR 75,77}]\).

(ii) the Norton theorem approach introduced by Chandy et al \([\text{CHAN 75}]\).

Both approaches give exact results for exponential networks, but several problems arise when applying the models to general queueing networks. In the following study these problems are enumerated and some solutions are suggested.

In section 7.2 a general survey is given. The problems are
 enumerated in section 7.3. The contribution of the maximum entropy analysis to the solution is given in section 7.4. Applications and examples are considered in section 7.5.

7.2 An Overview

In this section a short survey of the two main approaches to the decomposition of queueing networks is given. The near-complete-decomposability (NCD) approach is discussed in section 7.2.1, followed in section 7.2.2 by the Norton theorem (NT) approach.

7.2.1 The near-complete-decomposability (NCD) approach, Courtois [COUR 75, 77]

This approach is generally based on the fact that the joint state probability distribution of the number of customers in the network can be expressed as the solution of a finite continuous time Markov chain:

\[ p_{t+1}(n) = p_t(n)Q^* \]  

(7.1)

where \( Q \) is the state transition probability matrix with elements \( q^*_{ij} \) representing the transition probability from state \((n_0, n_2, \ldots, n_i, \ldots, n_M)\) to state \((n_0, \ldots, n-1, \ldots, n+1, \ldots, n_M)\), where \( n_i \) represents the number of customers at node \( i \), \( i=1, \ldots, M \), and \( \sum_{i=1}^{M} n_i = N \), \( N \) being the total number of customers in the network. At equilibrium, the time subscript \( t \) is dropped and 7.1 becomes:

\[ p(n) = p(n)Q^* \]  

(7.2)
Fig. 7.1 A central server model

Fig. 7.2 The equivalent two-server model for the central server network of Fig. 7.1
Since $Q^*$ is a stochastic matrix, then $p(n)$ can be interpreted as the left eigenvector of $Q^*$ corresponding to a maximal eigenvalue of unity.

As an example, the state transition probability matrix of the queueing network shown in Fig. 7.1 is given in Fig. 7.4. For convenience here the nodes are numbered from 0 to $M-1$. If in this matrix, the probabilities outside the solid-line squares (along the diagonal) are of lower magnitude than those inside the squares, then $Q^*$ is said to be nearly-completely decomposable, or diagonally dominant. Moreover, if inside each solid square, the elements outside the dotted squares are of lower magnitude than those on the inside, $Q$ is said to be 2-level nearly-completely-decomposable. In principle, this process can be continued to give an $(L+1)$-level nearly-completely-decomposable system, where $L=M-1$ and $M$ is the total number of nodes.

Based on the above definitions and the eigen-structures of the transition matrix $Q$, Courtois [COUR 75, 77] derived a criterion for the NCD of queueing networks, which can be stated as follows:

A stochastic matrix $Q$ is said to be $(L-1)$-level nearly-completely decomposable, if for each level $\ell$, $\ell=1, \ldots, L-1$, the following inequality is satisfied:

$$\omega_\ell < \frac{1}{2} (A_\ell + B_\ell) - (A_\ell B_\ell)^{\frac{1}{2}} \cos \left\{ \frac{\pi}{(N+1)} \right\}$$  \hspace{1cm} (7.3)

where

$$\omega_\ell = \max \left\{ \sum_{n=0}^{N} \sum_{k=0}^{L} \delta(n_k) \mu_k \sum_{m=0}^{L} P_{km} + \sum_{k=0}^{L} \delta(n_k) \mu_k \sum_{m=L+1}^{N} P_{km} \right\}$$  \hspace{1cm} (7.4)

$$A_\ell = \min_{0<k<\ell-1} (\mu_k P_{kl})$$  \hspace{1cm} (7.5)

$$B_\ell = \mu_\ell \sum_{k=0}^{\ell-1} P_{kl}$$  \hspace{1cm} (7.6)
Fig. 7.3 The procedure of multilevel aggregation
Fig. 7.4 The transition matrix $Q$ for the system in Fig. 7.1
Courtois further demonstrated that, if the above criterion is satisfied, then one can distinguish between short-term equilibrium probabilities and the required long-term equilibrium probabilities. The short-term equilibrium solutions can be obtained by the aggregation procedure illustrated in Fig. 7.3. Namely, the analysis is started at level 1 by considering the two-node network composed of node 0 and node 1 only. Following this, node 0 and node 1 are aggregated to form an equivalent composite queue. At level 2, the two-node network consisting of this composite queue and node 2 is analysed. Then a new composite queue is formed and the process is hierarchically continued until level L−1 where L=M−1 and M is the total number of nodes.

At each level \( \ell \), the input to the composite or aggregate node (Fig. 7.2) is taken to be

\[
\lambda_\ell = \mu_\ell \sum_{k=0}^{\ell} p_{\ell k}
\]

where \( p_{\ell k} \) is the probability that a customer leaving node \( \ell \) goes to node \( k \).

The output of the composite queue is obtained as the aggregation variable:

\[
R_{\ell,k}(n_{\ell}) = \{1 - \nu_{\ell}(n_{\ell}|n_{\ell})\} \mu_\ell \sum_{n_{\ell-1}}^{n_{\ell}} \nu_{\ell}(n_{\ell-1}|n_{\ell}) R_{\ell-1,k}(n_{\ell-1})
\]

(7.8)

\( \ell \geq 1, \ell = 1, \ldots, L-1; n_{\ell} = 1, \ldots, N; n_{\ell-1} = 1, \ldots, n_{\ell} \)

and \( \nu_{\ell}(n_{\ell-1}|n_{\ell}) \) is the short-term equilibrium probability that \( n_{\ell-1} \) customers among \( n_{\ell} \) customers are found in the composite (aggregate) node, which in turn can be obtained using the formula:
\[
\nu_{\ell}(n_{\ell-1}|n_{\ell}) = \frac{\sum_{k=0}^{\ell-1} \pi_{\ell} \prod_{k=1}^{\ell-1} R_{\ell-1,k}^{(k)} v_{\ell}^{(0)}(n_{\ell})}{\prod_{k=1}^{\ell-1} R_{\ell-1,k}^{(k)}} v_{\ell}^{(0)}(n_{\ell})
\]

(7.9)

At level 1, of course

\[
R_{o,k}(n_0) = \mu_0 p_{o,k} \text{ for all } n_0 \neq 0
\]

(7.10)

The aggregation procedure can therefore be summarised by the following algorithm:

**Algorithm 7.1a**

(i) For \( \ell = 1 \), compute \( R_{o,k}(n_0) \) using (7.10).

Compute \( \nu_1(n_0|n_1) \) using (7.9)

(ii) For \( \ell = 2, \ldots, L-1 \) compute \( R_{\ell,k}(n_{\ell}) \) using (7.8).

Compute \( \nu_{\ell}(n_{\ell-1}|n_{\ell}) \) using (7.9)

The philosophy behind the above aggregation procedure is that the low-level nodes attain their short-term equilibrium before the high-level nodes. Consequently, at each level \( \ell \), interaction with higher level \( \ell+1, \ldots, L-1 \) can be disregarded.

Once the short-term equilibrium probabilities \( \nu_{\ell}(n_{\ell-1}|n_{\ell}) \) have been obtained, the long-term equilibrium probabilities can be obtained by disaggregating \( \nu_{\ell}(n_{\ell-1}|n_{\ell}) \). To this end define \( p_{\ell}(n_{\ell}) \) to be the unconditional long-term equilibrium probability that there are \( n_{\ell} \) customers at the single node \( \ell \) and define \( b_{\ell}(n_{\ell-1}) \) to be the unconditional long-term equilibrium probability that there are \( n_{\ell-1} \) customers in the complimentary aggregate (composite) node. These probabilities and the performance metrics can be obtained using the following algorithm:
Algorithm 7.1b

(i) initialise:
\[ b_L(n_{L-1}) = v_L(n_{L-1}|N) \quad p_L(i_L) = p_L(N-i_L|N) \]  \hspace{1cm} (7.11)
\[ n_{L-1} = 0, 1, \ldots, N \quad i_L = 0, \ldots, N \]

(ii) For each level \( k = L-1, \ldots, 1 \)

compute
\[ b_{k}(n_{k-1}) = \sum_{n_k=n_{k-1}}^{N} b_{k+1}(n_k) v_k(n_k-1|n_k) \]  \hspace{1cm} (7.12)
\[ p_{k}(i_k) = \sum_{n_k=i_k}^{N} b_{k+1}(n_k) v_k(n_k-i_k|n_k) \]  \hspace{1cm} (7.13)

(iii) for each node \( \ell, \ell = 0, \ldots, M-1 \)

obtain

the utilisation \( \rho_{\ell} = 1-p_{\ell}(o) \) \hspace{1cm} (7.14)

the average number of customers
\[ <n>_{\ell} = \sum_{n_{\ell}}^{N} n_{\ell} p_{\ell}(n_{\ell}) \]  \hspace{1cm} (7.15)

It has been proved that the above algorithms give exact results for exponential networks [VANT 78] therefore they are equivalent to the Buzen algorithm. In fact the Buzen algorithm can be incorporated in the aggregation process to make it shorter, as hinted by Courtois himself [COUR 77]. Courtois also demonstrated that his algorithm can easily be generalised to work for networks with arbitrary service times. This generalisation will be discussed in section 7.4.
7.2.2 The Norton's theorem approach

This approach has been introduced by Chandy, Herzog and Woo \cite{CHAN75}. It is generally based on Norton's theorem for electric circuits which implies that the behaviour of a component in a given electric circuit can be studied by constructing an equivalent circuit in which all the other components are replaced by one composite component.

Analogously, Chandy et al deduced that the behaviour of a certain node $i$ in a given queueing network can be studied by constructing an equivalent network in which all the other nodes in the given network are replaced by a single equivalent composite queue. Therefore for each node $i, i=1, \ldots, M$, in the given network, there exists a reduced two-node network consisting of node $i$ and its "complementary composite node" $B_1$, Fig. 7.7, such that the equilibrium solution of the node $i$ in the equivalent network is identical to that in the given network. The parameters of the composite node $B_1$ are obtained from the original network by setting the service time of node $i$ equal to zero (replace it by a short). The service rate of $B_1$ can then be taken equal to the conditional throughput of the corresponding subnetwork $\text{SUB}Qi$, Fig. 7.6, $R_i(n)$, when there are $n$ customers in the subnetwork and $N-n$ customers at node $i$, the service time of node $i$ being zero. The whole procedure is illustrated by Figs. 7.5-7.7. Chandy et al proved that this method gives exact results for product-form networks and they developed an iterative algorithm to obtain the solution.

Later, Sauer and Chandy \cite{SAUE75} developed a heuristic flow-equivalent algorithm for central server models with general service times. The algorithm is based on Norton's theorem and can be described as follows.
Fig. 7.5 The original network (flow-equivalent method).

Fig. 7.6 The equivalent subnetwork SUBQi with node i replaced by a short.
Fig. 7.7 The equivalent two-server model
Algorithm 7.2

(i) Reduce the central server model to a two-node model consisting of the central server on one hand and a composite queue formed by aggregating the remaining nodes on the other hand.

(ii) Determine the output of the composite queue using the Buzen algorithm. The output of the composite queue is taken equal to the state dependent throughput of the network from which the central server is omitted (shorted) (Fig. 7.6).

(iii) Determine the coefficient of variation for the composite queue using the relation

\[ C_{sc}^2 = \sum_{k=1}^{M-1} C_{sk}^2 P_{ok} \]  \hspace{1cm} (7.16)

where \( M-1 \) is the number of nodes aggregated and \( o \) denotes the central server.

(iv) Select an exponential stage-representation for the composite queue depending on the coefficient of variation computed in (iii): if \( C_{sc}^2 < 1 \) then Erlang, if \( C_{sc}^2 > 1 \), then hyperexponential and if \( C_{sc}^2 = 1 \), then pure exponential.

(v) Solve the two-node network to obtain the performance metrics of the central server (Fig. 7.7).

(vi) Obtain the performance metrics of the individual queues constituting the composite queue.
7.3 Problems of Aggregation

As mentioned above, both near-complete decomposability and Norton's theorem lead to exact results when applied to exponential queueing networks. However several problems arise when these methods are applied to general queueing networks. These problems can be summarised as follows:

(i) Since the flow in general queueing networks is not Poisson, merging or splitting this flow leads to complicated flow processes that are not easy to determine. This problem is particularly apparent in the case of Norton's theorem approach and has been previously discussed by Sevick et al [SEVI 77].

(ii) In both approaches the analyst is faced with the analysis of two-server models that have general service times at both centres. The existing algorithms cannot easily cope with this problem.

(iii) In case of the flow-equivalent method, the throughput of the subnetwork $\text{SUBQ}_i$ of Fig. 7.6 is evaluated assuming exponential service times at the individual nodes. This assumption is heuristic and needs to be validated.

(iv) Although Courtour [COUR 77] has provided a criterion for near-complete decomposability and a measure for the error, no such measures exist for the methods based on Norton's theorem. In particular, the flow-equivalent algorithm outlined above is heuristic.

(v) Finally, it has often been noted that the above two approaches are very similar, yet no serious attempt has been made to merge them into a unified approach.
7.4 Contribution of the Maximum Entropy Analysis

Most of the problems mentioned in the above section have indirectly been addressed in the previous chapters e.g. the flow equations for the composite queue can easily be obtained by a slight modification of algorithm 6.1. Analysis of two-node models with general service times at both centres can be carried out using the methods developed in Chapter 5. In this section Courtois' algorithm and the flow-equivalent algorithm are accordingly modified to work for queueing networks with general service times.

7.4.1 Modification of the Courtois Algorithm

To modify Courtois' algorithm, the flow problem can be stated more clearly as follows:

In the two-node network of level \( \ell \) (which consists of a composite queue and node \( \ell \), see Fig. 7.3):

(i) Determine the service time squared coefficient of variation for the composite queue consisting of nodes 0,1,\ldots,\ell-1.

(ii) Determine the arrival process squared coefficient of variation for the composite queue.

(iii) Determine the output rate and the departure process squared coefficient of variation.

The service time distribution of the composite queue can be obtained using the merging procedure described in section 6.3.2. To this end assume that the service time distribution at each node is of the form (3.15) and define

\[
\alpha_{1\ell} = \sum_{k=0}^{\ell-1} p_{k\ell} l'_{k}\quad \text{and} \quad \alpha_{2\ell} = \sum_{k=0}^{\ell-1} p_{k\ell} l''_{k}
\]
Applying equations (6.21)-(6.24) successively one obtains:

\[ F^c_x(t) = 1 - \frac{a_{1x}}{a_{2x}} \exp\{-a_{1x}t\} \quad (7.17) \]

which is the distribution function for the composite queue service time. From this the pdf is:

\[ f^c_x(t) = (1 - \frac{a_{1x}}{a_{2x}})u_0(t) + ((a_{1x})^2/a_{2x}) \exp\{-a_{1x}t\} \quad (7.18) \]

Consequently the squared coefficient of variation for the composite service time is:

\[ C_{cs}^2 = \left(\frac{2a_{2x}}{a_{1x}}\right)^2 - 1 \quad (7.19) \]

From Fig. 7.2 the arrival process to the composite queue is nothing but the departure process from node \( x \). Similarly the departure process from the composite queue represents the arrival process to node \( x \). Therefore algorithm (6.1) can be slightly modified to compute the flow for the composite queue as follows:

**Algorithm 7.3a**

(i) For each level \( x \) compute the coefficient of variation of the composite queue using (7.19).

(ii) Repeat

a. Compute departure process coefficient of variation for the composite queue using (6.31).

b. Compute the \( C_{dx}^2 \) for node \( x \) using (6.31).

c. Set the arrival process coefficient of variation of the composite queue to \( C_{dx}^2 \) until \( C_{dc}^2 \) converges.
(iii) Compute
\[ \begin{align*}
\gamma_{ac} &= \frac{(C_{ac}^2 - 1)}{2}, \\
\gamma_{sc} &= \frac{(C_{sc}^2 - 1)}{2}
\end{align*} \]

The output of the composite queue can still be computed using (7.8), as shown by Courtois [COUR 77].

In view of this (7.9) can be modified to give

\[ \nu_{\lambda}(n_{k-1}|n_{\lambda}) = \nu_{\lambda}(0|n_{\lambda}) \prod_{k=1}^{n_{\lambda}-1} x_{kl} \]  \hspace{1cm} (7.20)

where

\[ x_{kl} = \frac{\Omega_{k\lambda} + Y_{ac} + \Omega_{k\lambda}Y_{sc}}{1 + Y_{ac} + \Omega_{k\lambda}Y_{sc}} \]  \hspace{1cm} (7.21)

and

\[ \Omega_{k\lambda} = \frac{\sum_{i=0}^{\lambda-1} a_{k\lambda} \prod_{i=0}^{k-1} p_{\lambda i}}{R_{\lambda-1,\lambda}(k)} \]  \hspace{1cm} (7.22)

Given this new formula for \( \nu_{\lambda}(n_{k-1}|n_{\lambda}) \) algorithm 7.1a and 7.1b can easily apply to obtain the required performance metrics. For an implementation, see the Appendix.

7.4.2 Modification of the flow-equivalent algorithm

There are two sources of error in the algorithm 7.2. First, in obtaining the flow parameters the departure process is ignored. Secondly, the equation used to compute \( C_{sc}^2 \) in step (iii) is heuristic.

A better estimation for the output of the composite queue can now be obtained using algorithm 6.1 with the Buzen algorithm. This procedure has already been applied in Chapter 6.
Similarly, $C_{sc}^{2}$ can now be computed, using (7.19). Step (iv) is no longer needed, and can be omitted.

7.5 Comparisons and Applications

7.5.1 Similarities between Courtois' algorithm and the flow-equivalent algorithm

From the previous sections the following points can be observed about Courtois' algorithm and the flow-equivalent algorithm:

1) The aggregation method is the same in both cases. The flow-equivalent algorithm corresponds to the step taken at level $L-1$ in algorithm 7.1a.

2) Equation (7.8) for determining $R_{k}^{(n,k)}$ can be replaced by step (ii) of the flow-equivalent algorithm 7.2, at level $(L-1)$ and vice versa without changing the value of $R_{k}^{(n,k)}$. I.e., the output of the composite queue can be obtained using either Buzen's algorithm or algorithm 7.1a.

3) The disaggregation method is the same in both cases.

Based on these points, it can be deduced that the flow-equivalent algorithm is a special case of the Courtois algorithm where all the levels $0,\ldots,L-2$ are omitted. The advantages of this relation can be summarised as follows:

a) The criterion for near-complete decomposability can now be used for the flow-equivalent method.

b) Step (ii) of the flow-equivalent algorithm for determining the output of the composite queue is now validated by theorem (6.1), p.79 of Cour 77 which implies that even in the case of arbitrary service times, the output of the composite queue is still given by (7.8), which
Fig. 7.8 Model of the example in section 7.5.2

Fig. 7.9 Model 2 of the example in section 7.5.2
Fig. 7.10 Model 3 of the example of section 7.5.2
Table 7.1 Performance metrics for the models of Figs. 7.8-7.10

<table>
<thead>
<tr>
<th>Mod. No.</th>
<th>Method</th>
<th>$Q_1$</th>
<th>$Q_2$</th>
<th>$Q_3$</th>
<th>$Q_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>utilis.</td>
<td>&lt;n&gt;</td>
<td>utilis.</td>
<td>&lt;n&gt;</td>
</tr>
<tr>
<td>1</td>
<td>Flow-equivalent</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SAUE 75a</td>
<td>0.59</td>
<td>0.91</td>
<td>0.59</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Buzen, Chap. 6</td>
<td>0.562</td>
<td>0.872</td>
<td>0.581</td>
<td>0.901</td>
</tr>
<tr>
<td></td>
<td>Courtois, modified</td>
<td>0.589</td>
<td>0.827</td>
<td>0.589</td>
<td>0.827</td>
</tr>
<tr>
<td>2</td>
<td>Flow-equivalent</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SAUE 75a</td>
<td>0.72</td>
<td>1.83</td>
<td>0.72</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Buzen, Chap. 6</td>
<td>0.762</td>
<td>2.0297</td>
<td>0.693</td>
<td>1.657</td>
</tr>
<tr>
<td></td>
<td>Courtois, modified</td>
<td>0.779</td>
<td>1.83</td>
<td>0.779</td>
<td>1.83</td>
</tr>
<tr>
<td>3</td>
<td>Flow-equivalent</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SAUE 75a</td>
<td>0.571</td>
<td>0.837</td>
<td>0.571</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Buzen, Chap. 6</td>
<td>0.544</td>
<td>0.815</td>
<td>0.558</td>
<td>0.828</td>
</tr>
<tr>
<td></td>
<td>Courtois, modified</td>
<td>0.551</td>
<td>0.761</td>
<td>0.551</td>
<td>0.761</td>
</tr>
</tbody>
</table>
is equivalent to step (ii) of algorithm 7.2.
c) The Buzen algorithm can be used to shorten the aggregation process in the Courtois algorithm.

7.5.2 Examples

Following the above discussion, the modified Courtois algorithm is applied to the models of Figs. 7.8-7.10. The results are compared with those of Sauer [SAUE 75a] and the modified Buzen algorithm of Chapter 6, (Table 7.1).

7.6 Summary

The maximum entropy results of the previous sections have been used to modify the Courtois algorithm and the flow-equivalent algorithm. The modified algorithms give satisfactory results, but more work is needed to investigate the effect of the flow parameters in aggregation. It has also been shown that the flow equivalent algorithm is a special case of the Courtois algorithm, which leads to unified aggregation procedure and removes many heuristics from the flow-equivalent algorithm.
CHAPTER VIII

CONCLUSIONS AND FUTURE PROSPECTS

8.1 General Summary

In this study maximum entropy has been used to analyse both single and multiple resource queueing models of computer system performance. In all cases, the form of distribution obtained is basically independent of any stochastic or operational assumptions and only requires the determination of the average number of customers in the system and the percentage of time the server is busy. If either the value or form of these parameters is provided, the corresponding maximum entropy distribution can directly be determined.

However, since in many prediction problems the expected number of customers in the system cannot be directly measured, it is often useful to approximate its actual value by a Markovian mean value formula. Good examples for these are the P-K mean formula (3.2) and the mean value formula (4.25), which are both independent of the pdf for the service or interarrival time and only depend on the first and second moments of these distributions. The only obstacle is that (4.25) requires the knowledge of the first and second moment of the idle time distribution, which can either be directly measured or estimated using the methodology of Chapter 4. Note that the entropy of a Markov process increases as this process evolves [KARL 81].

This combination of maximum entropy and Markov theory makes sense and solves a number of long-standing problems in the performance analysis of computer systems. The power of this approach has been illustrated
by applying the methodology to various types of models:

For the case where $C_a^2 = 1$ and $C_s^2$ is any finite, positive real number, a unique product-form solution, satisfying local balance has been derived for the first time. It has then been shown that this solution corresponds to the steady state solution of the underlying Markov process when the service time distribution is of the generalised exponential (GE) type (3.15). The generalised exponential is the sum of an exponential term and a unit impulse (generalised function) at the origin. The appearance of the unit impulse function in this distribution should not be surprising since it has been identified as the identity element in Kingman's algebra for queues [KLEI 75].

For the case where $C_s^2 = 1$ and $C_a^2$ is any finite, real positive number, the maximum entropy solution is of the same form as the steady state solution of the corresponding Markov process. However, it has been shown that a GE-type interarrival time distribution leads to a solution that maximises the entropy over all two-stage phase type distributions.

In the more general case where both $C_a^2$ and $C_s^2$ are arbitrary, real positive numbers, there have been three main achievements:

(i) Firstly, the spectral methods have been promoted to give analytic mean value formulae for different types of the G/G/1 system. This has been done by the effective use of Rouche's theorem. For the first time, analysts will be able to obtain the actual value of the average number of customers and the mean waiting time for a wide range of G/G/1 systems. Previously only approximations and simulation methods were used. The methodology developed has then been applied to study the effect of the service and interarrival time pdf on the performance metrics. The results
revealed that arbitrary assumptions about these functions may have a serious effect on the performance metrics. These metrics are expected to be invariant to the service time pdf only in the case of the M/G/l and the GE/G/l. Since, in many practical cases, the pdf of the service or interarrival time is arbitrarily fixed, the obtained results may unnecessarily be distorted.

(ii) To avoid this problem, maximum entropy has been suggested as a criterion. A maximum entropy model has been used to determine a form of solution for general single queues. It has then been explained that this form of solution is closely associated with that of the steady state solution for the G/GE/1 system. Following this, the maximum entropy solution for the G/GE/1 was determined and characterised. The implementation of the maximum entropy model in the operational (direct) sense was discussed and useful maximum entropy queue recursions were derived.

(iii) Thirdly, because of the special importance of the G/GE/1 and the GE/GE/1 in the analysis of general networks, the flow in these systems has been separately studied. The idle time distribution and the departure process for each system was characterised. The effect of feedback on the interinput distribution for the GE/GE/1 was also studied.

The results obtained for single general queues were then applied to analyse general two-server tandem and cyclic queues. Tandem queues were used to evaluate the effect of approximating the interdeparture time distribution by a GE-type distribution. The analysis of two-server cyclic models required the determination of the maximum entropy solution for the G/GE/1/N which has been obtained from that of the infinite capacity G/GE/1. The numerical results obtained for cyclic queues are compared with diffusion and simulation results.
For general queueing networks, it has been shown that maximum entropy produces a general product form solution that can either be implemented directly or in conjunction with Markovian queueing theory. Direct implementation of the solution requires the measurement of $<n>_i$ and $\rho_i$ which is possible if the model is to be used in consistency checking or performance calculation. If the model is to be implemented in the context of Markovian queueing theory, then the moments of the flow in the network must be computed. For this purpose, a new algorithm based on the GE-type distribution has been suggested. The methodology developed was then used to analyse examples of open and closed queueing networks. The results obtained compare favourably with diffusion, simulation and other results in the literature.

The ideas developed were further used to simplify and promote the aggregation techniques given previously by Chandy et al [CHAN 75] and Courtois [COUR 77].

8.2 Advantages of the Analysis

The advantages of the approach developed in this study can be summarised as follows:

(i) The analysis offers a reformulation of the queueing problem where $p_n$, the probability distribution of the number of customers in the system, is fixed irrespective of the pdf of the service or interarrival time. In this way the analyst is saved the trouble of making arbitrary and often distorting assumptions. The maximum entropy approach is similar in this sense to the diffusion and operational methods.

(ii) It is believed that the maximum entropy distribution is the one which is experimentally realised in overwhelmingly more ways than any

*Note that this algorithm does not take into consideration the split process described by Sevick et al [SEVI 77]. This may be the reason that the results of tables 6.4, 6.5 and 7.1 are inconsistent in comparison with the exact results given by Sevick et al [SEVI 77]. Further investigations are needed.
other distribution [JAYN 57a, 68]. This view has further been reinforced by the entropy concentration theorem [JAYN 83]. More than that, the maximum entropy distribution includes all the prior information about the model.

(iii) For the first time, a methodology has been described for determining the average number of customers and the mean waiting time in the G/G/1.

(iv) The class of locally balanced networks has been extended to include general queueing networks with FCFS disciplines.

(v) Maximum entropy greatly simplifies the analysis and removes the unnecessary complexity. It is interesting that the performance metrics of non-exponential queueing networks can still be obtained using Buzen's algorithm.

(vi) The analysis defines a new promising trend that can have wide implications on the performance evaluation of computer systems. The basic tools provided can now easily be used to obtain the response times, utilisations, throughputs, etc. of computer and communications networks, no matter what values are given for the service time coefficients of variation. With the increasing importance of distributed processing, the maximum entropy models will be exceedingly useful.

(vii) The maximum entropy model is a useful bridge between Markovian and operational methods.

8.3 Future Prospects

Despite the above gains, the present study is in many ways incomplete. However, the approach given is promising and can be easily extended and promoted. To achieve this, the following points should be
observed:

1. A first priority is to give proofs for the few conjectures used at various stages of the study.

2. It should be investigated if the first and second moment of the idle time distribution can be obtained without involving the pdfs of the service and interarrival time distributions. This will make formula 4.25 and consequently the maximum entropy distribution $p_n$ more independent of these functions.

3. The effect of serial correlation in the interinput process on the mean waiting time in the G/G/1 should be further investigated, based on the results obtained in this study. An attempt has been made in this respect by Jenkins [Jenk 66] and Shimshak [Shim 81]. If this effect is serious the Markovian processes should be replaced by time series.

4. The similarity between maximum entropy and the diffusion approximation should be investigated and exploited. One way of achieving this is to add a variance constraint to the maximum entropy model of Chapter 4 and see how the maximum entropy solution is affected.

5. The flow algorithm described in Chapter 6 can quite possibly be promoted to give better results. It should also be investigated if it is better to compute the flow moments in this way or to assume a Poisson interinput process for each queue in the network.

6. The significance and properties of the GE-type distribution (3.15) should be further investigated.

7. The role of the entropy functional in the decomposition of general queueing networks should be investigated, see [Vane 69].

8. The methodology given in Chapter 4 can easily be extended to give useful results for the G/GE/M. However, it may be very difficult to
obtain actual parameter values beyond the G/GE/2.

9. The implications of Lemma 4.1 should be further investigated.

10. The impact of maximum entropy on the simulation of general queueing networks should be investigated. In this respect, there are two potential applications: random number generation and determining confidence intervals. It will be interesting to see the effect of the GE-type distribution in generating random numbers.

To conclude, the maximum entropy formalism has proved itself as a satisfactory and convincing theory for general queueing networks. It does not generally contradict the existing Markovian queueing theory results but extends and refines them. More research is needed to promote the methodology and the underlying techniques.
REFERENCES


BASK 75 Baskett, F; Chandy, K M; Muntz, R R & Palacios, 'Open, closed and mixed networks of queues with different classes of customers', JACM, 22, 248-260, (1975).


ELAF 83  El-Affendi, M A & Kouvelatos, D D: 'A maximum entropy analysis of the M/G/1 and G/M/1 queueing systems at equilibrium'. To appear in Acta Informatica (1983).

ELAF 83  El-Affendi, M A & Kouvelatos, D D: 'Spectral methods and maximum entropy in the analysis of the G/G/1', submitted to the JACM.


