1

37535

National Library
of CanadaBibliothèque nationale
du CanadaCANADIAN THESES
ON MICROFICHETHÈSES CANADIENNES
SUR MICROFICHENAME OF AUTHOR/NOM DE L'AUTEUR Elhamy E. AbdouTITLE OF THESIS/TITRE DE LA THÈSE DIFFUSION APPROXIMATION OF SOME MULTICLASS QUEUEING SYSTEMS
(WITH APPLICATION TO FRONT-END AND HOST INTERFACE CONFIGURATIONS)UNIVERSITY/UNIVERSITÉ CARLETON UNIVERSITYDEGREE FOR WHICH THESIS WAS PRESENTED/
GRADE POUR LEQUEL CETTE THÈSE FUT PRÉSENTÉE Ph.D.YEAR THIS DEGREE CONFERRED/ANNÉE D'OBTENTION DE CE DEGRÉ 1978NAME OF SUPERVISOR/NOM DU DIRECTEUR DE THÈSE Dr. Samy A. Mahmoud

Permission is hereby granted to the NATIONAL LIBRARY OF
CANADA to microfilm this thesis and to lend or sell copies
of the film.

The author reserves other publication rights, and neither the
thesis nor extensive extracts from it may be printed or other-
wise reproduced without the author's written permission.

*L'autorisation est, par la présente, accordée à la BIBLIOTHÈ-
QUE NATIONALE DU CANADA de microfilmer cette thèse et
de prêter ou de vendre des exemplaires du film.*

*L'auteur se réserve les autres droits de publication; ni la
thèse ni de longs extraits de celle-ci ne doivent être imprimés
ou autrement reproduits sans l'autorisation écrite de l'auteur.*

DATED/DATE April 26th, 1978. SIGNED/SIGNÉ E. AbdouPERMANENT ADDRESS/RÉSIDENCE FIXE 1218 Meadowlands Drive,Apartment #308,Ottawa, Ontario,

National Library of Canada

Cataloguing Branch
Canadian Theses Division

Ottawa, Canada
K1A 0N4

Bibliothèque nationale du Canada

Direction du catalogage
Division des thèses canadiennes

NOTICE

AVIS

The quality of this microfiche is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us a poor photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this film is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30. Please read the authorization forms which accompany this thesis.

**THIS DISSERTATION
HAS BEEN MICROFILMED
EXACTLY AS RECEIVED**

La qualité de cette microfiche dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de mauvaise qualité.

Les documents qui font déjà l'objet d'un droit d'auteur (articles de revue, examens publiés, etc.) ne sont pas microfilmés.

La reproduction, même partielle, de ce microfilm est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30. Veuillez prendre connaissance des formules d'autorisation qui accompagnent cette thèse.

**LA THÈSE A ÉTÉ
MICROFILMÉE TELLE QUE
NOUS L'AVONS REÇUE**

The undersigned recommend to the Faculty of Graduate Studies acceptance
of the thesis:

DIFFUSION APPROXIMATION OF SOME MULTICLASS QUEUEING SYSTEMS
(WITH APPLICATION TO FRONT-END AND HOST INTERFACE CONFIGURATIONS)

submitted by Elhamy E. Abdou, B.Sc.(Hon.), M.Eng., in partial fulfilment
of the requirements for the degree of Doctor of Philosophy.

S. Mahmoud
.....
Thesis Supervisor

Bernard Pagurek
.....
Thesis Supervisor

David A. Coll
.....
Chairman,
Department of Systems Engineering
and Computing Science

Carleton University

April 6, 1978

DIFFUSION APPROXIMATION OF SOME MULTICLASS QUEUEING SYSTEMS
(WITH APPLICATION TO FRONT-END AND HOST INTERFACE CONFIGURATIONS)

by

Elhamy E. Abdou

A thesis submitted to the Faculty of
Graduate Studies in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy

Faculty of Engineering
Carleton University
Ottawa, Ontario
February 1978

© 1978 Elhamy E. Abdou

ABSTRACT

This research presents satisfactory approximations to the steady state solution of some practical queueing systems and networks with two classes of customers. Such approximations are needed when simplifying assumptions such as the Poisson arrival and the exponential service time are inadequate or invalid.

Diffusion approximations are developed for GI/G/1 queueing systems with two classes of customers and the following service disciplines: (1) Preemptive-Resume, (2) First-come-first-served (FCFS) with different service time distributions for the different classes, and (3) Alternating priority. A rather extensive study of the accuracy of the approximations for the previous disciplines is included. For each discipline, the approximate solutions and the analytical or simulation solutions are compared in terms of the means, the variances, and the marginal distributions of the number of customers for each class in the system. The accuracy of the diffusion approximations is found to be quite satisfactory in most cases. For the FCFS and the alternating priority disciplines, the accuracy is considerably higher for symmetrical Markov queueing systems (Poisson arrival, exponential service times, equal rates of arrival for both classes, and equal rates of service for both classes). Empirical expressions have been proposed to improve the accuracy of other non-symmetrical systems.

In the area of open queueing networks, we have extended Reiser and Kobayashi's algorithm [RK74] to networks with two classes of customers and the service disciplines mentioned before. Examples are given to

show the application of the new algorithm. Estimates of the mean number of customers of each class at every center, and estimates of the total delay in the network are examined and compared with simulation. The accuracy is very satisfactory and superior to the Poisson assumption in estimating queues with non-Poisson arrival.

For unbalanced closed networks with two classes of customers, we have proposed a method to obtain an equivalent open network for the unsaturated part which can be solved by the previous algorithm. Examples are presented and comparison of the approximations and the simulation results have shown satisfactory accuracy.

We have developed a model for the communication subsystem in a teleprocessing system. The model describes the behavior of a Front-End Processor connected to a Host computer via a channel. It has the form of an open queueing network. The model for the service facility in both processors permits different ways of function partitioning between the two processors. The diffusion approximation algorithm for open networks is used in an example to compare between the performance of four configurations with different function partitioning. We have also proposed a "cost-effectiveness" measure for more practical comparison between these configurations.

ACKNOWLEDGEMENTS

I would like to thank Dr. B. Pagurek for his guidance and many helpful suggestions throughout the course of this thesis. In particular, I appreciated his constructive criticism during the writing.

Thanks are due also to Dr. S. Mahmoud for his constant encouragement and help during the course of this research. The author is grateful to Dr. Mahmoud for suggesting the problem of the performance analysis for the communication subsystem.

I cannot find words to adequately express my appreciation to my family, especially my mother, for their support and unshakeable faith in my ability to complete this endeavour.

I am extremely grateful to my friends for their assistance and support. Special thanks to Mr. M.B. Khalil for his help in drawing the figures.

The skilful typing of Mrs. R. Thomas is highly appreciated.

This work was financially supported by the National Research Council of Canada.

TABLE OF CONTENTS

	PAGE
CHAPTER 1 - INTRODUCTION	1
1.1 Introduction	1
1.2 Scope and Motivation	1
1.3 Review of Related Work	5
1.4 Thesis Contributions	6
1.5 Thesis Organization	8
CHAPTER 2 - PRELIMINARIES	11
2.1 Introduction	11
2.2 The One-Dimensional Diffusion Process	12
2.2.1 Definition and Properties	12
2.2.2 The Diffusion Equation	13
2.2.3 Boundary Conditions	14
2.2.4 Transient and Steady State Solutions	17
2.3 The Multidimensional Diffusion Process	19
2.4 Diffusion Approximation for Single Class Queueing Systems ..	22
2.4.1 Introduction	22
2.4.2 Diffusion Analysis of the Number of Customers in the Queueing System	23
CHAPTER 3 - DIFFUSION APPROXIMATION FOR PREEMPTIVE-RESUME QUEUEING SYSTEMS	27
3.1 Introduction	27
3.2 The Departure Process	28
3.3 The Diffusion Equation and the Steady State Solutions	35
3.4 Analysis of Results	41
3.4.1 Introduction	41
3.4.2 Accuracy of the Diffusion Approximation for M/G/1 Queueing Systems	41
3.4.3 Accuracy of the Diffusion Approximation for GI/G/1 Queueing Systems	46
3.5 Concluding Remarks	55
CHAPTER 4 - DIFFUSION APPROXIMATION FOR AN FCFS SINGLE SERVER WITH TWO CLASSES	59
4.1 Introduction	59
4.2 The Departure Process	59
4.3 The Diffusion Equation	64
4.4 Boundary Conditions and Steady State Solutions	68

TABLE OF CONTENTS (Continued)

	PAGE
4.5 Analysis of Results	80
4.5.1 Introduction	80
4.5.2 Accuracy of the Diffusion Approximation for M/M/1 Systems	80
4.5.3 Accuracy of the Diffusion Approximation for GI/G/1 Systems	88
4.6 Concluding Remarks	119
CHAPTER 5 - DIFFUSION APPROXIMATION FOR ALTERNATING PRIORITY QUEUEING SYSTEMS	121
5.1 Introduction	121
5.2 The Departure Process	123
5.3 The Diffusion Equation and the Steady State Solutions	128
5.3.1 The Diffusion Equation	128
5.3.2 Steady State Solutions	129
5.4 Analysis of Results	133
5.4.1 Introduction	133
5.4.2 Accuracy of the Diffusion Approximation for M/G/1 Queueing Systems	133
5.4.3 Accuracy of the Diffusion Approximation for GI/M/1 Queueing Systems	146
5.5 Concluding Remarks	159
CHAPTER 6 - NETWORKS OF QUEUES	161
6.1 Introduction	161
6.2 Review of Literature	161
6.2.1 Exact Solutions for Open Networks	161
6.2.2 Approximate Solutions for Open Networks	164
6.3 Open Networks with More Than One Class of Customers	171
6.3.1 Introduction	171
6.3.2 The Arrival Processes	172
6.3.3 The Algorithm	176
6.4 Numerical Results for Open Networks	180
6.4.1 Introduction	180
6.4.2 Example 1	180
6.4.3 Example 2	189
6.5 Diffusion Analysis of Closed Networks with Two Classes of Customers	196
6.5.1 Review of Literature on Closed Networks	196
6.5.2 Closed Networks with Two Classes	201
6.5.3 The Algorithm	204

TABLE OF CONTENTS (Continued)

	PAGE
6.6 Numerical Results for Closed Networks	205
6.6.1 Introduction	205
6.6.2 Example 3	206
6.6.3 Example 4	212
6.7 Concluding Remarks	216
CHAPTER 7 - DESIGN OF FRONT-END AND HOST PROCESSOR INTERFACE CONFIGURATIONS	217
7.1 Introduction	217
7.2 A Review of the Approaches Used in Analysing Communication Processors	219
7.3 Modelling the Communication Subsystem	221
7.4 A Communication Subsystem Example	234
7.4.1 Performance Analysis Study	234
7.4.2 Cost-Effectiveness Study	237
7.5 Concluding Remarks	240
CHAPTER 8 - CONCLUSIONS AND SUGGESTIONS FOR FUTURE WORK	243
8.1 Conclusions	243
8.2 Recommendations for Future Work	245
REFERENCES	246
APPENDIX A	251
APPENDIX B	255
APPENDIX C	260

LIST OF TABLES

TABLE		PAGE
3.1	Effect of Changing c_2 on the Accuracy for $\alpha = .5$, $\beta = .125$, $c_1 = 1$ and $\rho_{\text{eff}} = .9$	44
3.2	Effect of Changing c_2 on the Accuracy for $\alpha = .3$, $\beta = .292$, $c_1 = 1$ and $\rho_{\text{eff}} = .9$	44
3.3	Effect of Changing ρ_{eff} (or β) on the Accuracy for $\alpha = .3$, $\rho_1 = .2$ and $c_1 = c_2 = 1$	45
3.4	Effect of Changing α or β on the Accuracy for $\rho_1 = .1$, $\rho_2 = .8$, $c_1 = 1$ and $c_2 = 1$	45
3.5	Accuracy of the Diffusion Approximation for Case 1	52
3.6	Accuracy of the Diffusion Approximation for Case 2	53
3.7	Accuracy of the Diffusion Approximation for Case 3	56
3.8	Accuracy of the Diffusion Approximation for Case 4	57
4.1	Accuracy of the Diffusion Approximation for Case 1	94
4.2	Accuracy of the Diffusion Approximation for Case 2	96
4.3	Accuracy of the Diffusion Approximation for Case 3	98
4.4	Accuracy of the Diffusion Approximation for Case 4	100
4.5	Accuracy of the Diffusion Approximation for Case 5	102
4.6	Accuracy of the Diffusion Approximation for Case 6	107
4.7	Accuracy of the Diffusion Approximation for Case 7	109
4.8	Accuracy of the Diffusion Approximation for Case 8	112
4.9	Accuracy of the Diffusion Approximation for Case 9	114
4.10	Accuracy of the Diffusion Approximation for Case 10	116
5.1	Accuracy of the Diffusion Approximation for Case 1	151
5.2	Accuracy of the Diffusion Approximation for Case 2	153
5.3	Accuracy of the Diffusion Approximation for Case 3	155
5.4	Accuracy of the Diffusion Approximation for Case 4	157

LIST OF TABLES (Continued)

TABLE		PAGE
6.1	Accuracy vs. Utilization of Server 2	186
6.2	Results of Example 2 for Open Networks	195
6.3	Accuracy vs. Throughput for Example 2	197
6.4	Accuracy of Diffusion Approximation for Example 3	210
6.5	Accuracy vs. $c_{1,2}$ for Example 3	212
6.6	Accuracy of Diffusion Approximation for Example 4	215
7.1	Telecommunication Functions	234
7.2	Mean Service Times for FEP and Host in 4 Different Configurations	235
7.3	Comparison between 4 FEP-Host Interface Configurations .	240
A.1	Table of Definitions	252
B.1	Table of Definitions	256
C.1	Table of Definitions	261

LIST OF FIGURES .

FIGURE .		PAGE
1.1	Communication Subsystem Model	4
3.1(a)	Example of a Time-Diagram for Departures	29
3.1(b)	Equivalent Time-Diagram for Departures	29
3.2	h-Stage Service Facility	49
3.3	Distribution of the Number in the System of the Low Priority Customers for Case 1	54
3.4	Distribution of the Number in the System of the Low Priority Customers for Case 2	54
3.5	Distribution of the Number in the System of the Low Priority Customers for Case 3	58
3.6	Distribution of the Number in the System of the Low Priority Customers for Case 4	58
4.1	Contour Lines	70
4.2	Relative Error (ER) as a Function of Correlation Coefficient r for Two Cases	77
4.3	Correlation Coefficient that Minimizes the Relative Error ER as a Function of $c_{av}/c_{d_{av}}$	79
4.4	Relative Error (ER) as a Function of α for $\beta = 1$ and Different Values of ρ_{eff}	82
4.5	Relative Error (ER) as a Function of α for $\rho_{eff} = 0.9$ and Different Values of β	82
4.6	Relative Error in the Mean Number in the System of Class 1 as a Function of α for $\beta = 1$ and Different Values of ρ_{eff}	84
4.7	Relative Error in the Mean Number in the System of Class 1 as a Function of β for $\alpha = 0.5$ and Different Values of ρ_{eff}	85
4.8	Relative Error in the Mean Number in the System of Class 1 as a Function of β for $\rho_{eff} = 0.95$ and Different Values of α	86

LIST OF FIGURES (Continued)

	PAGE
4.9 Relative Error in the Mean Number in the System of Class 1 as a Function of α for $\rho_{eff} = 0.95$ and Different Values of β	87
4.10 Comparison Between the Relative Error ER for $r = -1$ and r^* ($\rho_{eff} = 0.9$)	89
4.11 Comparison Between the Relative Error in Estimates of Class 1 for $r = -1$ and r^* ($\rho_{eff} = 0.95$)	90
4.12 Marginal Queue Length Distributions for Case 1	95
4.13 Marginal Queue Length Distributions for Case 2	97
4.14 Marginal Queue Length Distributions for Case 3	99
4.15 Marginal Queue Length Distributions for Case 4	101
4.16 Marginal Queue Length Distributions for Case 5	103
4.17 Accuracy of Diffusion Approximation for the Mean Number in the System of Class 1 as a Function of β for Different Values of ρ_{eff} , with $\alpha = 0.5$, $c_1 = c_2 = 1.583$, and $r = -1$	104
4.18 Effect of α on the Accuracy of the Diffusion Approximation of the Mean Number in the System of Class 1 Customers for $\beta = 1$, $\rho_{eff} = 0.91$, and $r = -1$.	105
4.19 Marginal Queue Length Distributions for Case 6	108
4.20 Marginal Queue Length Distributions for Case 7	110
4.21 Effect of Changing the Squared Coefficient of Variation of Service Time on the Accuracy of Estimating the Mean Number in the System of Class 1 for $\rho_{eff} = 0.91$, $\alpha = 0.5$, and $\beta = 1$	111
4.22 Marginal Queue Length Distributions for Case 8	113
4.23 Marginal Queue Length Distributions for Case 9	115
4.24 Marginal Queue Length Distributions for Case 10	117

LIST OF FIGURES (Continued)

	PAGE
4.25 Effect of Changing the Squared Coefficient of Variation of Arrival Times on the Accuracy of Estimating the Mean Number in the System of Class 1 for $\rho_{\text{eff}} = 0.91$, $\alpha = 0.5$, and $\beta = 1$	118
5.1 Time-Diagram Example for the Alternating Priority Service Discipline	126
5.2 Correlation Coefficient that Minimizes the Relative Error ER as a Function of the Queueing System Parameters	132
5.3 Accuracy vs. ρ_{eff} for $\beta = 1$ and $c_{a1} = c_{a2} = c_1 = c_2 = 1$	136
5.4 Accuracy vs. β for $\alpha = 0.5$ and $c_{a1} = c_{a2} = c_1 = c_2 = 1$	136
5.5 Accuracy vs. c_s for $\alpha = 0.5$, $\rho_{\text{eff}} = 0.9$ and $c_{a1} = c_{a2} = 1$	137
5.6 Accuracy vs. c_s for $\alpha = 0.5$, $\beta = 2$, and $c_{a1} = c_{a2} = 1$	137
5.7 ER_1 vs. α for $\beta = 1$, and $c_a = c_s = 1$	138
5.8 ER_1 vs. β for $\alpha = 0.5$, and $c_a = c_s = 1$	140
5.9 ER_1 vs. β for $\alpha = 0.4$, and $c_a = c_s = 1$	141
5.10 ER_2 vs. β for $\alpha = 0.4$ and $c_a = c_s = 1$	142
5.11 ER_1 vs. c_s for $\alpha = 0.5$, $\beta = 1$, and $c_a = 1$	143
5.12 ER_1 vs. c_s for $\alpha = 0.5$, and $c_a = 1$	144
5.13 ER_1 vs. c_s for $\alpha = 0.4$, $\beta = 1$, and $c_a = 1$	145
5.14 ER vs. c_s for $\alpha = 0.5$, $\rho_{\text{eff}} = 0.9$, $c_{a1} = c_{a2} = 1$, and Two Values for r	147
5.15 ER_1 vs. β for $\alpha = 0.5$, $c_a = c_s = 1$, and Two Values for r	148

LIST OF FIGURES (Continued)

	PAGE
5.16 ER_1 vs. c_s for $\alpha = 0.5$, $\rho_{eff} = 0.98$, $c_{a1} = c_{a2} = 1$ and Two Values for r	149
5.17 Marginal Queue Length Distributions for Case 1	152
5.18 Marginal Queue Length Distributions for Case 2	154
5.19 Marginal Queue Length Distributions for Case 3	156
5.20 Marginal Queue Length Distributions for Case 4	158
5.21 Accuracy vs. c_a for $\alpha = 0.44$, $\beta = 1$, and $c_s = 1$	160
6.1 Open Network of Example 1	181
6.2 Delay $T_{1,1,3,2}$ vs. Utilization of Server 2	187
6.3 Delay $T_{3,1,1,2}$ vs. Utilization of Server 2	188
6.4 Open Network of Example 2	190
6.5 Delay $T_{1,1,3,2}$ vs. Throughput	198
6.6 Delay $T_{3,1,1,2}$ vs. Throughput	199
6.7 Cyclic Network of Example 3	207
6.8 The Distribution of the Number of Customers in Center 1 of Example 3	211
6.9 Central Server Model of Example 4	213
6.10 Reduced Model	213
7.1 Teleprocessing System	223
7.2 Communication Subsystem	223
7.3 FEP Model	226
7.4 Service Time Model	226
7.5 Models for Channel	229

LIST OF FIGURES (Continued)

		PAGE
7.6	Software Organization for a Teleprocessing System	229
7.7	Host CPU	231
7.8	Model for Disk	231
7.9	Communication Subsystem Model	233
7.10	Communication Delay as a Function of Throughput for 4 Different Configurations of Front-End and Host Processors Interface	238
7.11	Cost-Effectiveness for 4 Different Configurations of Communication Subsystems	241
A.1	GPSS Block Diagram for a Preemptive-Resume Queueing System with Two Classes of Customers	253
B.1	GPSS Block Diagram for an FCFS Queueing System with Two Classes of Customers	258
C.1	GPSS Block Diagram for an Alternating Priority Queueing System with Two Classes of Customers	263

LIST OF SYMBOLS

The most frequently used symbols are listed below:

$A_i(t)$	The Cumulative Number of Arrivals of Class i Customers up to Time t
a_{ij}	The Infinitesimal (i,j) Covariance
b_i	The Infinitesimal i^{th} Mean
c_{a_i}	The Squared Coefficient of Variation for the Interarrival Times of the i^{th} Class
$c_{a_i,l}$	The Squared Coefficient of Variation for the Interarrival Times of Class l at Center i
c_{d_i}	The Squared Coefficient of Variation for the Interdeparture Times of the i^{th} Class
$c_{d_i,l}$	The Squared Coefficient of Variation for the Interdeparture Times of Class l at Center i
c_i	The Squared Coefficient of Variation for the Service Times of the i^{th} Class
$c_{i,l}$	The Squared Coefficient of Variation for the Service Times of Class l at Center i
$D_i(t)$	The Cumulative Number of Departures of Class i Customers up to Time t
ER_i	The Relative Error in the Diffusion Approximation of the Mean Number of Customers of Class i in the System
ERV_i	The Relative Error in the Diffusion Approximation of the Variance of the Number of Customers of Class i in the System
$Q_i(t)$	The Number of Class i Customers in the Queueing System (Including the One in Service, if any) at Time t
N	Number of Service Centers in a Network
n	Number of Customers in a Queueing System

n_i	Number of Customers of Class i in a Queueing System
$\hat{n}_i$	Diffusion Approximation of n_i
$n_{i,l}$	Number of Class i Customers in Center i
$\hat{n}_{i,l}$	Diffusion Approximation of $n_{i,l}$
r	The "infinitesimal" Correlation Coefficient of the Normally Distributed Variables $\Delta Q_1(t)$ and $\Delta Q_2(t)$
$r_{i,l;j,m}$	The Probability that a Customer of Class i at Center i will Join Class m Customers at Center j , upon Completion of Service at Center i
$X_i(t)$	The Continuous Stochastic Process that Approximates the Discrete Process $Q_i(t)$
α	The Probability of Next Arrival being of Class 1
β	The Ratio of the Mean Service Rate of Class 2 to the Mean Service Rate of Class 1
$\Delta A_i(t)$	The Number of Arrivals of Class i Customers between Time t and $t + dt$
$\Delta D_i(t)$	The Number of Departures of Class i Customers between Time t and $t + dt$
$\Delta Q_i(t)$	The Change in the Number of Customers of Class i in the System between Time t and $t + dt$
λ_i	The Mean Rate of Arrival of Class i Customers
$\lambda_{i,l}$	The Mean Rate of Arrival of Class i Customers at Center i
μ_i	The Mean Service Rate of Class i Customers
$\mu_{i,l}$	The Mean Service Rate of Class i Customers at Center i
μ_{di}	The Mean Rate of Departure of Class i Customers
$\mu_{d,i,l}$	The Mean Rate of Departure of Class i Customers at Center i
ρ_{eff}	Total Utilization of the Server
ρ_i^*	Total Utilization of Server i

LIST OF ABBREVIATIONS

i.i.d.	Independently and Identically Distributed
PDF	Probability (or Cumulative) Distribution Function
pdf	Probability Density Function
r.v.	Random Variable
var/cov	Variance/Covariance
M/M/1	Single Server Queue with Poisson Arrival and Exponential Service Times
M/G/1	Single Server Queue with Poisson Arrival and General Service Times
GI/G/1	Single Server Queue with Independent General Interarrival Times and General Service Times
FCFS	First-Come-First-Served
I/O	Input-Output
FEP	Front-End Processor

1. INTRODUCTION

1.1 Introduction

Queueing systems' models for which analytical solutions have been developed are limited in their use since they do not represent complex practical problems. Very often, it is necessary to introduce approximations to make such systems analytically tractable. Examples of these approximations are the Poisson arrival and the exponential service time assumptions. Furthermore, the available analytical results for general models are often expressed in forms so complex as to make them of little help in answering questions of practical nature. To find answers to such practical questions researchers and system designers turn to simulation techniques. However, complex computer simulations are costly in terms of their design and their computer running time. The cost of simulation also increases rapidly as the number of parameters of the queueing system increases.

This thesis addresses the problem of finding approximate solutions for certain practical queueing systems and networks. The objective is to make these solutions simple, yet accurate to a degree that makes them useful for system designers.

1.2 Scope and Motivation

One of the practical problems that motivated the development of the approximate solutions reported in this thesis is the design of the communication subsystem in modern teleprocessing networks. This problem is discussed briefly here, and the details are presented in Chapter 7.

In teleprocessing systems, the communication subsystem performs some of the telecommunication functions (such as assembling, routing, buffering, queueing, and editing) on messages before handing them to the appropriate terminal or application program. With the increase in the power of mini-computers and microprocessors together with the rapid decrease in their cost, it was apparent that dedicating a small processor to carry out the telecommunication functions will lead to offloading the Host processor and improving the system cost/performance. However, many design problems have not been answered. These problems concern the choice of the hardware configuration, and the partitioning of the telecommunication functions between the two processors.

In chapter 7, a model is developed for the communication subsystem. The model is for a Front-End processor connected to a Host computer via a half-duplex channel (e.g. an IBM 370X communication processor connected to an IBM 370 system). Each of the two processors is modelled by single server handling two priority tasks with a Preemptive-Resume discipline. The high priority tasks represent the I/O transfer of messages. The output from this level goes to the low priority queue to wait for further processing.

The service facility for the low priority tasks is modelled as an N -stage server, where stage i has exponentially distributed service time with mean rate μ_i . The probability that a message will request service from stage i is π_i and the probability that the message will skip this stage is $(1 - \pi_i)$. This model for the service facility enables us to change the division of the telecommunication functions between the two processors.

The channel is modelled as a single server with two classes of customers and alternating priority discipline. The mass storage device (e.g. disk)

used for queueing messages is modelled by a first-come-first-served (FCFS) queueing system with general service time, and one or two classes of customers according to the number of processors who can access the storage device.

Therefore, the FEP-Host model is a queueing network as shown in fig. 1.1, in which servers have external (Preemptive-Resume) or internal (FCFS, and alternating) priority disciplines.

Some progress has been reported in the area of analytical solutions for queueing networks [JJ57, GN67, Ba75, Ch77]. However, Chandy et al [Ch77] showed that to have the product form analytical solution all service centers in the network must satisfy at least one of the following conditions:

- (1) service to new customers must begin immediately on their arrival (e.g. Last-come-first-served, Infinite server, or Processor sharing), and
- (2) service time distributions are negative exponential with the same mean service rate for all classes.

It is obvious that our FEP-Host model does not satisfy the previous two conditions. Therefore, approximate solutions were considered.

Recently, a technique based on the diffusion approximation was developed to provide approximate solutions for queueing systems and networks with one class of customers and FCFS queueing discipline [Ga68, Ne71, Ko74, Ge75]. The technique was applied to solve queueing systems under heavy traffic conditions and results with satisfactory accuracy were reported [RK74]. In addition to satisfactory accuracy, diffusion approximation solutions are relatively inexpensive to use compared to iterative approximate techniques.

Fig. 1.1 Communication Subsystem Model

In this thesis, we pursue further the development of diffusion approximation techniques for solving queueing systems and networks with two (or more) classes of customers, while considering several service disciplines such as Preemptive-Resume, FCFS, and alternating priority.

1.3 Review of Related Work

Several advances have been made recently in the development of diffusion approximation techniques for single servers with one class of customers and FCFS service discipline.

Newell [Ne68, Ne71] used the diffusion approximation to study queues with time-dependent arrival rate. His study concerns the transition of M/G/1 systems through saturation, or in other words the "rush hour situation".

Gaver [Ga68] applied the diffusion approximation to the waiting time in an M/G/1 queueing system, and obtained the transient as well as the steady state virtual load. Fischer [Fi77] extended the steady state results to the case of loop service system where the server cyclically serves a fixed number of peripheral stations.

Kobayashi [Ko74] obtained the steady state distribution for the number of customers in a GI/G/1 queueing system. Gelenbe [Ge75] improved the previous solution by considering a different set of boundary conditions for the diffusion equations. Heyman [He75] obtained an approximation for the time-dependent distribution of queue length and virtual delay in a GI/G/1 queueing system using diffusion approximations.

In the area of diffusion approximation for queueing systems with multiservers, Newell [Ne73] obtained results for GI/G/m with large values

of m , i.e. $m \geq 1$, while Halachimi and Franta [HF77] presented an approximate solution which yields uniformly good results for all values of m .

Diffusion approximations were also used successfully for open and closed networks with FCFS queueing centers and one class of customers. Kobayashi [Ko74] obtained solutions for open and closed networks using the multidimensional diffusion equation and a set of reflecting boundary conditions. Reiser and Kobayashi [RK74] suggested simpler algorithms that solve for the marginal distributions first and then obtain the joint distribution from the product form. These algorithms require considerably less computation than those reported in [Ko74], with some sacrifice in the accuracy of the results. Gelenbe and Pujolle [GP76] improved the previous algorithms by calculating more accurate values for the squared coefficient of variation of the interdeparture times.

Other work in the area of networks includes:

- (1) the diffusion approximation of CPU utilization in a cyclic queue with general service times and FCFS queueing discipline by Gaver and Shedler [GS73];
- (2) the diffusion approximation of the repairman problem with two types of failure [GL77], and
- (3) the analysis of tandem queues with application to the special case of 2 queues and equal service rates [Ne75].

1.4 Thesis Contributions

The major contributions of this thesis are summarized in the following:

1. Developing diffusion approximation for the distribution of the number of customers in a GI/G/1 queueing system with multi(two)-class

customers and Preemptive-Resume, FCFS, or alternating priority service disciplines (Chapters 3, 4 and 5).

2. Developing algorithms for the solution of open and closed networks with multi(two)-class customers and the previous service disciplines. For closed networks the algorithm provides solutions for unbalanced networks (Chapter 6).

3. Developing a queueing network model for a communication subsystem composed of two processors connected via a channel. The diffusion approximation algorithm for open networks is then used for analyzing the performance of the subsystem (Chapter 7).

Several other contributions have resulted from this research:

1. The derivation of expressions for the distribution of the inter-departure times of each class in GI/G/1 queueing systems with two classes of customers and Preemptive-Resume, FCFS, or alternating priority service disciplines (Chapters 3, 4 and 5).

2. The introduction of a new h-stage service facility and obtaining expressions for the first two moments of the total service time (Chapter 3).

3. Obtaining empirical expressions for the "infinitesimal" correlation coefficient between the r.v.^s representing the changes (in time dt) in the number of customers of each class in FCFS, and alternating priority queueing systems (Chapters 4 and 5).

4. Obtaining expressions for the squared coefficient of variation of interarrival times to any center in multi(two)-class queueing networks (Chapter 6).

5. Proposing procedures to detect bottlenecks in unbalanced closed queueing networks and to transform the unsaturated part of the closed

network to an open network (Chapter 6).

6. Proposing "cost-effectiveness" measure for studying computer communication networks and using this measure for comparing communication subsystem configurations (Chapter 7).

1.5 Thesis Organization

This thesis is organized as follows:

Chapter 2 gives an overview of the one dimensional diffusion process and the multidimensional diffusion process, and the corresponding diffusion equations. Different boundary conditions are discussed briefly, and transient and steady state solutions of some cases are presented. Finally, the main results of the diffusion approximation for a single queue with FCFS service discipline are reported.

Chapter 3 is devoted to the study of the diffusion approximation for a single server queueing system with two classes of customers and Preemptive-Resume service discipline. First, expressions are obtained for the distribution and moments of the interdeparture times of the low priority class in an M/G/1 system. These results are applicable in the limit for GI/G/1 queueing systems. Second, a two dimensional diffusion equation is obtained together with its steady state solution, under reflecting boundary conditions. This solution gives the approximate marginal distribution for the number of customers of the low priority class in the queueing system. Third, the accuracy of the approximations is examined by comparison with analytical formulas and/or simulation.

In addition, a new h-stage service facility is introduced in this chapter, and expressions for the first two moments of the service time are

obtained. This special service facility is used in some of the numerical examples.

Chapter 4 is concerned with the diffusion approximation of a GI/G/1 queueing system with FCFS service discipline. Expressions for the distribution and moments of the interdeparture times of each class in an M/G/1 system are derived. These expressions apply in the limit to GI/G/1 systems. Then, the diffusion equation and its steady state solution are obtained. The results of the approximation are compared with analytical expressions and/or simulation, and an empirical expression for the "infinitesimal" correlation coefficient is obtained which results in accuracy improvements.

Chapter 5 presents the diffusion approximation of a GI/G/1 queueing system with two classes of customers and alternating priority. First, the departure processes of the two classes are approximated by renewal processes, and expressions for the distribution of their interdeparture times are obtained. Second, the diffusion equation, and its steady state solutions are derived. Third, the accuracy of the results are examined by comparison with analytical expressions and/or simulation, and an empirical expression to improve the accuracy is obtained.

Chapter 6 presents an algorithm for the solution of open networks with two classes of customers and the previously discussed service disciplines. Expressions for the squared coefficient of variation of interarrival times of each class at each center in the network are derived. Unbalanced closed networks are also studied, and an algorithm is presented to detect bottlenecks in the network and transform the unsaturated part to an open network. Some examples of open and closed networks are presented. The

accuracy of the approximation, as a function of some of the network parameters, is examined.

Chapter 7 is devoted to the development of the communication subsystem model discussed before. The diffusion approximation algorithm for open networks (developed in chapter 6) is then used to study the performance of a number of configurations which differ in telecommunication functions partitioning between the FEP and the Host Processor. A cost-effectiveness measure is proposed for practical comparison between the configurations.

Chapter 8 contains some concluding remarks and recommendations for future work.

2. PRELIMINARIES

2.1 Introduction

In this chapter two diffusion processes are introduced and discussed briefly; these are the one-dimensional diffusion process and the multidimensional diffusion process. The use of the one-dimensional diffusion process as an approximation for queue behavior is also reviewed.

In sec. 2.2.1 the one-dimensional diffusion process is defined and its properties are discussed. The diffusion equation that describes this stochastic process is presented in sec. 2.2.2, and a number of boundary conditions, associated with the one-dimensional diffusion equation, are discussed in sec. 2.2.3. The transient and steady state solutions of the diffusion equation with one reflecting boundary conditions are presented in sec. 2.2.4.

Section 2.3 describes the multidimensional diffusion process. The multidimensional diffusion equation is presented first, followed by the special case of the two-dimensional diffusion equation. Finally the reflecting boundary conditions are discussed using the notion of flux.

In sec. 2.4.1, the diffusion approximation of a queueing system with one class of customers and FCFS service discipline is reviewed. The main results of the diffusion approximation for the steady state distribution of the number of customers in the queueing system are presented in sec. 2.4.2.

2.2 The One-Dimensional Diffusion Process

2.2.1 Definition and Properties

The Diffusion Process is defined as a Markov Process in which only continuous changes of state occur. Thus it can be regarded as a continuous time - continuous state stochastic Markov process.

As a result of this definition, we find the following properties:

a) In a small time interval, a diffusion process can only undergo a small change of state. This is called the "Local character".

b) Because of the Markov Property, the diffusion process satisfies the Chapman-Kolmogorov equation, which relates the distribution function of the process at time t with that at initial time t_0 through the distribution at an intermediate time τ ,

$$P(x_0, t_0; x, t) = \int_{-\infty}^{\infty} P(y, \tau; x, t) d_y P(x_0, t_0; y, \tau), \quad t_0 < \tau < t \quad (2.1)$$

where $P(x_0, t_0; x, t)$ is the conditional distribution function of $X(t)$, assuming $X(t_0) = x_0$.

The corresponding equation for conditional probability densities (if they exist) is

$$p(x_0, t_0; x, t) = \int_{-\infty}^{\infty} p(x_0, t_0; y, \tau) p(y, \tau; x, t) dy, \quad t_0 < \tau < t \quad (2.2)$$

The Wiener Process is a special case of the Diffusion Process in which $X(t)$ has an additional property of having independent increments. Because of its importance, we state the definition of the Wiener process with drift 'b' and variance 'a' as a stochastic process $X(t)$ with the

following properties:

- a) $X(0) = 0$
- b) $\{X(t), t \geq 0\}$ has stationary independent increments.
- c) $X(t_2) - X(t_1)$ is normally distributed with mean $b(t_2 - t_1)$ and variance $a(t_2 - t_1)$ for any $t_1 < t_2$.

2.2.2 The Diffusion Equation

Consider the stochastic variable $X(t)$, for which the increment $dX(t)$ in a small time interval dt ($X(t + dt) - X(t)$) has mean $b(x, t)dt$ and variance $a(x, t)dt$ and is independent of all previous increments. Thus, $X(t)$ is defined by the stochastic differential equation

$$dX(t) = b(x, t)dt + (a(x, t))^{1/2} dW(t) \quad (2.3)$$

where $W(t)$ is a Wiener process with zero mean and unit variance.

From this definition, it can be shown that the conditional probability density function at time t , $p(x_0, t_0; x, t)$, given that $X(t_0) = x_0$, satisfies the following partial differential equation [CM65]

$$\frac{\partial}{\partial t} p(x, t) = \frac{1}{2} \frac{\partial^2}{\partial x^2} \{a(x, t)p(x, t)\} + \frac{\partial}{\partial x} \{b(x, t)p(x, t)\} \quad (2.4)$$

where $p(x, t) \triangleq p(x_0, t_0; x, t)$

This equation is known as the Fokker-Plank diffusion equation. It is also called the Kolmogorov forward equation. We will refer to it as the one-dimensional diffusion equation.

The quantities $b(x,t)$ and $a(x,t)$ are known as the infinitesimal mean and variance respectively, and are defined as

$$b(x,t) \triangleq \lim_{dt \rightarrow 0} \frac{E\{[X(t+dt) - X(t)] | X(t) = x\}}{dt} \quad (2.5)$$

$$\text{and } a(x,t) \triangleq \lim_{dt \rightarrow 0} \frac{\text{Var}\{[X(t+dt) - X(t)] | X(t) = x\}}{dt} \quad (2.6)$$

$$\text{where } \text{Var}[dX(t) | X(t) = x] \triangleq E\{[dX(t) - E[dX(t) | X(t) = x]]^2 | X(t) = x\}$$

These quantities give the rate of change of the conditional mean and variance at time t .

Another approach of deriving the one-dimensional diffusion equation is shown in Kleinrock [Kl76]. This approach starts from the Chapman-Kolmogorov equation, and uses Taylor expansion.

A third and interesting way to arrive at the same one-dimensional diffusion equation is shown in Cox and Miller [CM65]. They derive the equation as the continuous limit of a general random walk with variable transition probabilities.

2.2.3 Boundary Conditions

The solution of the one-dimensional diffusion equation of the previous section, given the initial conditions $X(0) = x_0$, will lead to an unrestricted process in which $-\infty \leq x(t) \leq \infty$ for all $t \geq 0$. It is obvious that such a process cannot be used to describe any physical process where the value of $x(t)$ should lie within a given finite range $x_1 \leq x(t) \leq x_2$ for $t \geq 0$.

For example, we may like to restrict $x(t)$ to positive values only, i.e. $x_1 = 0$ and $x_2 = \infty$. In order to achieve that, the solution has to satisfy additional equations that describe the behavior of the stochastic process when it reaches the boundaries. These equations are known as the boundary conditions.

Gittman and Skorohod [GS72] give some boundary conditions which are associated with the one-dimensional diffusion equation, and can be given physical interpretation. The following is a brief discussion of these boundary conditions.

1) Absorbing boundaries

We assume that the diffusion process $X(t)$ represents the position of a particle at time t with the initial condition x_0 . An absorbing boundary is placed at $x_1 > x_0$. It follows that the particle moves from its starting position until it reaches the boundary. Once there, it is no longer subject to perturbation and thus stays there forever. Cox and Miller [CM65] showed that the appropriate condition for an absorbing boundary at x_1 is that $p(x, t)$ must vanish at $x = x_1$ for all t , i.e.,

$$p(x, t) \Big|_{x=x_1} = 0, \quad t > 0$$

2) Reflecting boundaries

In this case, the particle reenters the interior of the space after hitting the boundary. We can differentiate between two types of reflection; namely, the instantaneous reflection and the delayed reflection.

In the instantaneous reflection case, the particle's exit velocity

from the boundary is infinite, and the time spent at the boundary can only be equal to zero. Otherwise an infinite displacement would occur in a finite time. Thus, this boundary can be considered completely elastic and the velocity of arrival to the boundary is the same as the exit velocity (which is infinite in the case of diffusion process).

In the delayed reflection, the particle reenters the interior of the space with finite velocity and positive time is required in order to attain finite displacement. The time that the particle spends on the boundary can be random with a given distribution (e.g. exponential). Such a process is still a Markov Process but does not follow the diffusion equation on the boundaries [CM65, Ge75]. For delayed reflection the velocity of the particle is partially damped and the boundary is said to be elastic.

Cox and Miller [CM65] showed that the boundary condition for instantaneous reflection at x_1 is

$$\left[\frac{1}{2} \frac{\partial}{\partial x} \{a(x,t)p(x,t)\} - b(x,t)p(x,t) \right]_{x=x_1} = 0 \quad (2.7)$$

Another way of looking at the instantaneous reflection is shown by Kimura [Ki64]. He introduces the notion of "flux" or the rate of net flow of probability across the point x , and proves that it is given by

$$\psi(x,t) = -\frac{1}{2} \frac{\partial}{\partial x} \{a(x,t)p(x,t)\} + b(x,t)p(x,t) \quad (2.8)$$

which satisfies the equation

$$\frac{\partial p(x,t)}{\partial t} = - \frac{\partial \psi(x,t)}{\partial x} \quad (2.9)$$

From this definition, it is easy to describe the instantaneous reflection as the one with zero flux across the boundary, and obtain the condition given by Eq. 2.7.

3) Jump-type boundaries

In this type, a particle vanishes after reaching the boundary and is replaced by a new one at an interior point x which starts a new diffusion.

It is possible to combine boundary conditions of the types discussed here to obtain more complicated boundaries.

The reflecting boundaries (instantaneous and delayed) are the most appropriate ones for queueing systems. In this thesis, we shall use the instantaneous reflecting boundaries to restrict queue lengths from becoming negative as will be shown later.

2.2.4 Transient and Steady State Solutions

In this section, we discuss solutions for the diffusion equation with constant coefficients; namely

$$\frac{\partial p(x,t)}{\partial t} = \frac{1}{2} a \frac{\partial^2}{\partial x^2} \{p(x,t)\} + b \frac{\partial p(x,t)}{\partial t} \quad (2.10)$$

where b and a are constants representing the infinitesimal mean and variance, respectively.

As we indicated before, this equation represents the behavior of a Wiener process with drift b and variance a . The unrestricted solution for $X(t)$ depends only on the initial condition. If we start the process at

time zero and fixed position x_0 , then $p(x_0, 0)$ is a delta function at x_0 . It is easy to show that $X(t)$ has the normal density function with mean $x_0 + b \cdot t$ and variance $a \cdot t$ given by [CM65]

$$p(x, t) = \frac{1}{\sqrt{2\pi at}} \exp \left\{ -\frac{(x - x_0 - bt)^2}{2at} \right\} \quad (2.11)$$

The solution for the restricted problem depends on the boundary conditions.

The solution for the diffusion equation with instantaneous reflection at the origin ($x(t) \geq 0$) is obtained by the method of images [CM65, Ne71, Ko74] as

$$P(x, t) = \frac{x - x_0 - bt}{(at)^{1/2}} - \exp \{2bx/a\} \frac{-x - x_0 - bt}{(at)^{1/2}} \quad (2.12)$$

where $P(x, t)$ is the distribution function, i.e.,

$$P(x, t) = \int_0^x p(y, t) dy$$

and $\Phi(x)$ is the standard normal integral;

$$\Phi(x) = (2\pi)^{-1/2} \int_{-\infty}^x \exp \{-z^2/2\} dz$$

The steady state or equilibrium distribution can be obtained by letting $t \rightarrow \infty$ in the transient solution. An alternative way is to solve

the diffusion equation with $\frac{\partial p(x,t)}{\partial t}$ set to zero. It is easy to show that for $b < 0$ the steady state solution is given as [CM65, Ko74]:

$$p(x) \stackrel{\Delta}{=} p(x,\infty) = (2|b|/a) \exp\{-2|b|x/a\} \quad (2.13)$$

whereas for $b \geq 0$, $p(x,t) \rightarrow 0$ as $t \rightarrow \infty$.

Gelenbe [Ge75] obtained the steady state distribution for a diffusion process with two delayed-reflection boundaries.

2.3 The Multidimensional Diffusion Process

Consider an n -dimensional vector $\bar{X}(t) \equiv \{X_i(t), i = 1, 2, \dots, n\}$ of diffusion processes with infinitesimal means defined as

$$b_i(\bar{x}, t) = \lim_{dt \rightarrow 0} \frac{1}{dt} E[\{X_i(t+dt) - X_i(t)\} | \bar{X}(t) = \bar{x}] \quad , \quad i=1, 2, \dots, n \quad (2.14)$$

where $\bar{x} \stackrel{\Delta}{=} \{x_i, i=1, \dots, n\}$

The infinitesimal covariances are defined as

$$a_{ij}(\bar{x}, t) = \lim_{dt \rightarrow 0} \frac{1}{dt} \text{cov}[\{X_i(t+dt) - X_i(t)\}, \{X_j(t+dt) - X_j(t)\} | \bar{X}(t) = \bar{x}] \quad (2.15)$$

The Fokker-Planck diffusion equation, then takes the form

$$\frac{\partial p(\bar{x}, t)}{\partial t} = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \frac{\partial^2}{\partial x_i \partial x_j} \{a_{ij}(\bar{x}, t) p(\bar{x}, t)\} - \sum_{i=1}^n \frac{\partial}{\partial x_i} b_i(\bar{x}, t) p(\bar{x}, t) \quad (2.16)$$

where $p(\bar{x}, t) \triangleq p(\bar{x}_0, t_0; \bar{x}, t)$ is the multivariate pdf of $\bar{X}(t)$ conditional on $\bar{X}(t_0) = \bar{x}_0 \triangleq \{x_{0i}, i = 1, \dots, n\}$

The corresponding stochastic differential equation is

$$d\bar{X}(t) = b(\bar{x}, t) dt + (a(\bar{x}, t))^{1/2} dW(t) \quad (2.17)$$

where $b(\bar{x}, t)$ is the vector $\{b_i(\bar{x}, t)\}$

$a(\bar{x}, t)$ is the symmetric positive definite matrix $\{a_{ij}(\bar{x}, t)\}$

$W(t)$ is an n -dimensional Wiener process with zero mean and unit covariance matrix.

For the case where the increments $d\bar{X}(t)$ given by

$$d\bar{X}(t) = \bar{X}(t + dt) - \bar{X}(t)$$

are mutually independent and have multivariate normal distribution with mean $b \cdot dt$ and covariance matrix $a \cdot dt$, the conditional joint density $p(\bar{x}, t)$ will satisfy the following diffusion equation with constant coefficients

$$\frac{\partial p(\bar{x}, t)}{\partial t} = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n a_{ij} \frac{\partial^2 p(\bar{x}, t)}{\partial x_i \partial x_j} - \sum_{i=1}^n b_i \frac{\partial p(\bar{x}, t)}{\partial x_i} \quad (2.18)$$

The notion of flux, introduced before, was extended to the two-dimensional diffusion equations by Kimura [K164]. In the remainder of

this section, we present his notion and relate the results to hydrodynamics.

Let $\psi_1\{x_1|x_2; t\}$ and $\psi_2\{x_2|x_1; t\}$ be the fluxes at a point (x_1, x_2) along the x_1 and x_2 axes given by

$$\psi_1 \triangleq \psi_1\{x_1|x_2; t\} = -\frac{1}{2} a_{11} \frac{\partial p(\bar{x}, t)}{\partial x_1} - \frac{1}{2} a_{12} \frac{\partial p(\bar{x}, t)}{\partial x_2} + b_1 p(\bar{x}, t) \quad (2.19)$$

$$\psi_2 \triangleq \psi_2\{x_2|x_1; t\} = -\frac{1}{2} a_{12} \frac{\partial p(\bar{x}, t)}{\partial x_1} - \frac{1}{2} a_{22} \frac{\partial p(\bar{x}, t)}{\partial x_2} + b_2 p(\bar{x}, t) \quad (2.20)$$

Then, the diffusion equation is expressed in the form

$$\frac{\partial p(\bar{x}, t)}{\partial t} = - \frac{\partial \psi_1}{\partial x_1} - \frac{\partial \psi_2}{\partial x_2} \quad (2.21)$$

$$\text{or} \quad - \frac{\partial p(\bar{x}, t)}{\partial t} = \nabla \cdot \psi$$

$$\text{where} \quad \nabla = \bar{i} \frac{\partial}{\partial x_1} + \bar{j} \frac{\partial}{\partial x_2}$$

$$\text{and} \quad \psi = \bar{i} \psi_1 + \bar{j} \psi_2$$

The orthogonal unit vectors $\bar{i}$ and $\bar{j}$ define a two-dimensional euclidean space.

This equation is known as the equation of continuity in hydrodynamics [Pi58]. The term $\nabla \cdot \psi$ is the divergence operation and represents the excess of the outward over the inward flow.

For incompressible flow, i.e. no divergence, the following condition must be satisfied

$$\nabla \cdot \psi = - \frac{\partial p(\bar{x}, t)}{\partial t} = 0 \quad (2.22)$$

For the two-dimensional case, an instantaneous reflecting boundary at $x_1 = 0$ can be imposed by letting the flux in the direction x_1 to be equal to zero as $x_1 \rightarrow 0$

i.e.

$$\psi_1\{x_1|x_2; t\}\Big|_{x_1=0} = 0 \quad (2.23)$$

Similarly an instantaneous reflecting boundary at $x_2 = 0$ is

$$\psi_2\{x_2|x_1; t\}\Big|_{x_2=0} = 0 \quad (2.24)$$

The steady state solution for this case will be discussed in the next chapter.

2.4 Diffusion Approximation for Single Class Queueing Systems

2.4.1 Introduction

Some of the characteristics of queueing systems display the local character discussed in sec. 2.2.1, i.e., small change of the characteristic will occur in a small time interval. An example of these characteristics is the queue length. To illustrate the local character property, let the queue length be $Q(t)$ at time t , then it is possible to choose a coarse scale for measurement of queue length such that the change in time dt of queue length will be an arbitrary small quantity $\Delta Q(t)$.

Another characteristic of queues that displays the local character is the virtual load, $v(t)$, defined as the total service time required to clear the queue at any time t . Since both queue length and virtual load are discrete-state processes, it is convenient to replace them by some corresponding continuous processes which can be described by differential

equations. The solution of these differential equations with an appropriate set of initial and boundary conditions will describe the transient and steady-state behavior of the original processes.

Approximations of the number of customers in the queueing system by diffusion processes can be found in Newell [Ne71], Kobayashi [Ko74], Heyman [He75], and Gelenbe [Ge75], and are discussed briefly in sec. 2.4.2. Related work was also developed by Gaver [Ga68] in which he presented a diffusion analysis for the transient as well as the steady virtual load of an M/G/1 queue. Heyman [He75] extends the results to GI/G/1 queues, and Fischer [Fi77] develops the steady state solution for the case of loop service system where the server cyclically serves a fixed number of peripheral stations.

2.4.2 Diffusion Analysis of the Number of Customers in the Queueing System

For a single server queueing system with one class of customer and FCFS queueing discipline, assume that $A(t)$ and $D(t)$ represent the cumulative number of arrivals and departures up to time t . Then, the number of customers in the queueing system (including the one in service) at time t is given by

$$Q(t) = A(t) - D(t)$$

The change in queue length between the time t and $t+dt$ is

$$\begin{aligned} \Delta Q(t) &= Q(t+dt) - Q(t) \\ &= \Delta A(t) - \Delta D(t) \end{aligned} \tag{2.25}$$

where $\Delta A(t) = A(t+dt) - A(t)$

$\Delta D(t) = D(t+dt) - D(t)$

If we replace the discrete process $Q(t)$ by a continuous one $X(t)$, then $X(t)$ will satisfy the diffusion equation of (Eq.2.10), if the arrival and service processes are stationary, and the queueing system is under heavy traffic.

To use this equation, we need the values of the infinitesimal mean b and the infinitesimal variance a . The approach taken by Newell [Ne71], Heyman [He75], and Kobayashi [Ko74] to obtain these parameters is to use the renewal theory results and the central-limit theorem. In this approach, assume the arrival and the departure processes are renewal with the mean and variance for interevent times given by $(\lambda^{-1}, \sigma_a^2)$ and (μ_d^{-1}, σ_d^2) respectively. Let dt be sufficiently large for many arrivals and departures to take place, and assume $Q(t)$ does not go to zero in the small interval dt (i.e. heavy traffic), then $\Delta Q(t)$ is approximately normally distributed with mean and variance given by,

$$E[\Delta Q(t)] = (\lambda - \mu_d)dt \quad (2.26)$$

$$\text{and} \quad \text{var}[\Delta Q(t)] = (c_a \lambda + c_d \mu_d)dt \quad (2.27)$$

$$\text{where} \quad c_a = \sigma_a^2 \lambda^2$$

$$c_d = \sigma_d^2 \mu_d^2$$

This result is based on a central-limit theorem of renewal processes and is discussed in detail in Cox and Miller [CM65] and Kleinrock [K176]. From these results and the definitions of the infinitesimal mean and

variance given by Eqs. 2.5 and 2.6, we find that:

$$b = (\lambda - \mu_d) \quad (2.28)$$

$$a = (c_a \lambda + c_d \mu_d) \quad (2.29)$$

Notice that because of the stationarity of the arrival and service processes, the infinitesimal mean and variance are not conditional.

It remains to calculate the distribution of the interdeparture times. Because of the heavy traffic assumption, the queue is never empty and the interdeparture times are equal to the service times. Furthermore, due to heavy traffic, we can say that $Q(t) > 0$ for all t . Thus, we can assume an instantaneous reflecting boundary is placed at $Q(t) = 0$ or $X(t) = 0$.

The solution of the diffusion equation approximating the queue length is similar to the one discussed in section 2.2.4 with the proper parameters. Thus, the steady state solution $p(x, \infty)$ is given as

$$p(x, \infty) = \frac{2(1-\rho)}{c_s + c_a \rho} \exp \left\{ \frac{-2(1-\rho)x}{c_s + c_a \rho} \right\}, \quad x \geq 0 \quad (2.30)$$

where $\rho = \lambda/\mu_d$ and $c_s = c_d$.

The approximate queue size distribution is obtained by discretizing the continuous solution as follows [Ko74]

$$\hat{p}(n) = \int_n^{n+1} p(x, \infty) dx = (1 - \hat{\rho}) \hat{\rho}^n \quad n=0,1,2,\dots \quad (2.31)$$

$$\text{where } \hat{\rho} = \exp \{ -2(1 - \rho)/(c_s + c_a \rho) \} \quad (2.32)$$

Kobayashi [Ko74] modified this result by taking into consideration the fact that

$$\hat{p}(0) = 1 - \rho$$

Thus,
$$\hat{p}(n) = \begin{cases} 1 - \rho & \text{for } n = 0 \\ \rho(1-\hat{p})\hat{p}^{n-1} & n \geq 1 \end{cases} \quad (2.33)$$

Gelenbe [Ge75] considered a slightly different approach by choosing delayed reflecting boundaries and arriving at a more formal way for deriving $\hat{p}(0)$ and some improvements in the results.

Another result in the diffusion approximations of the single queue is found in the analysis by Halachimi and Pranta [HP77] for the G/G/k systems.

3. DIFFUSION APPROXIMATION FOR PREEMPTIVE-RESUME QUEUEING SYSTEMS

3.1 Introduction

In some practical queueing systems, we wish to give some customers higher priority over other customers, and thus serve them ahead of those waiting in the queue with lower priority. This queueing discipline has the advantage of reducing the length of the waiting queue if the high priority is given to the customers who need small service time. Thus, it is suitable for applications in which customers can be divided into classes because their service time requirements are substantially different.

In this chapter, we discuss a priority scheme known as the Preemptive-Resume. In this scheme, the arrival of a customer with high priority interrupts the service of a customer with lower priority. The interrupted customer resumes service towards completion only after all the higher priority customers leave the system (i.e. no high priority customers are waiting in the queue or are being served).

The low priority customers do not interfere with the high priority class. In fact, the high priority customers see a completely dedicated server for their use. Therefore, the diffusion approximation of the queueing behavior of the high priority is exactly the one for a GI/G/1 system with one class discussed in Chapter 2.

In this chapter, we derive the diffusion equation for the low priority class. In sec. 3.2, we obtain an expression for the distribution of the interdeparture times. The diffusion equation and the steady state solution are derived in sec. 3.3. In sec. 3.4, we compare the diffusion results with analytical formulas and simulation runs.

3.2 The Departure Process

Assume λ_1 and λ_2 to represent the mean arrival rate of the high priority class, and the low priority class, respectively. Arrivals of each class are independent and Poisson, and hence the total arrival to the queue is likewise Poisson with mean rate $\lambda = \lambda_1 + \lambda_2$.

The service time for class i ($i = 1$ or 2) has a general cumulative distribution function $F_i(x_i)$ with mean μ_i^{-1} , variance σ_i^2 , and squared coefficient of variation c_i ($c_i = \sigma_i^2 \mu_i^2$). The ratio of the mean service rates is $\beta = \mu_2/\mu_1$. The probability of next arrival being of the high priority class is α and the probability of next arrival being of the low priority class is $(1 - \alpha)$, where $\alpha = \lambda_1/\lambda$.

As mentioned before, the preemptive policy implies that there will be no departures of the low priority class as long as there are customers of the high priority class in the system. We proceed now to present a new approach for studying the interdeparture times of the low priority class.

Let Z_2 be a r.v. representing the interdeparture time of the low priority class. Assume that we observed a departure of class 2 and we are waiting for the subsequent departure of that class. Since the system is assumed to run under heavy traffic, and the queue does not have any customers of class 1 at the instant of departure of class 2, the server will start serving the first customer in the queue of class 2. However, this can get interrupted m times ($0 \leq m \leq \infty$) if one or more customers of class 1 request service. The interrupted customers resume receiving service from the point of their interruption. Figure 3.1(a) shows a time diagram for the server for a case where there are 3 departures

Fig. 3.1(a) Example of a Time-Diagram for Departures

Fig. 3.1(b) Equivalent Time-Diagram for Departures

of class 1 customers before one customer of class 2 is fully serviced. This customer was interrupted twice. It is clear that fig. 3.1(b) is equivalent to fig. 3.1(a) in the interdeparture time of class 2 customers, and the number of departures of class 1 customers. In fig. 3.1(b), we lumped the service times given to the one class 2 customer and made it follow the three departure times of class 1 customers. In general, the interdeparture times of class 2 are composed of n service times of class 1 customers ($0 \leq n \leq \infty$) followed by one service time of class 2 customers.

Since these departure sequences are mutually exclusive events, we can use the law of total probability to calculate the PDF of the r.v. Z_2 as

$$\begin{aligned}
 P[Z_2 \leq z_2] &= P[Z_2 \leq z_2 \text{ and } \{2\}] + P[Z_2 \leq z_2 \text{ and } \{1,2\}] \\
 &\quad + P[Z_2 \leq z_2 \text{ and } \{1,1,2\}] + \dots \\
 P[Z_2 \leq z_2] &= P[X_2 \leq z_2 \text{ and } \{2\}] + P[X_1 + X_2 \leq z_2 \text{ and } \\
 &\quad \{1,2\}] + P[X_1^1 + X_1^2 + X_2 \leq z_2 \text{ and } \{1,1,2\}] \\
 &\quad + \dots
 \end{aligned} \tag{3.1}$$

where $\{X_1^1, X_1^2, X_1^3, \dots\}$ are independent non-negative r.v.^s

all with the same PDF $F_1(x_1)$, and X_2 is the r.v. with PDF $F_2(x_2)$.

The sequences of departure are represented by $\{2\}$, $\{1,2\}$, etc.

Therefore, $P\{2\}$ = Probability of a departure of class 2

= Probability of no arrivals of class 1
during service of class 2

$$= 1 - \alpha$$

$P\{1,2\}$ = Probability of the sequence; 1 followed by 2

= Probability of only one arrival of class 1
during service of class 2

$$= \alpha(1 - \alpha)$$

$P\{1,1,2\}$ = Probability of the sequence; 2 follows two
departures of 1

= Probability of two arrivals of class 1
during service of class 2

$$= \alpha^2 (1 - \alpha)$$

etc.

Also from the definition of the r.v.^s X_1 and X_2

$$P[X_2 \leq z_2] = F_2(z_2)$$

$$P[X_1^1 + X_2 \leq z_2] = F_1(z_2) \otimes F_2(z_2)$$

$$P[X_1^1 + X_1^2 + X_2 \leq z_2] = F_1(z_2) \otimes F_1(z_2) \otimes F_2(z_2) \\ = \{F_1(z_2)\}^2 \otimes F_2(z_2)$$

etc.

where $\otimes$ indicates the convolution operation defined as

$$F_1(z_2) \otimes F_2(z_2) \triangleq \int_{-\infty}^{\infty} F_1(z_2 - x_2) dF_2(x_2)$$

and $\{F_1(z_2)\}^n$ is an n-fold convolution of $F_1(z_2)$.

Since the sequence of departures is independent of service times, then

$$P[Z_2 \leq z_2] = (1 - \alpha)F_2(z_2) + \alpha(1 - \alpha) F_1(z_2) \otimes F_2(z_2) \\ + \alpha^2(1 - \alpha) \{F_1(z_2)\}^2 \otimes F_2(z_2) + \dots \quad (3.2)$$

and $H_2(z_2) \triangleq P[Z_2 \leq z_2]$

$$= (1 - \alpha) \sum_{n=0}^{\infty} \alpha^n \{F_1(z_2)\}^n \otimes F_2(z_2) \quad (3.3)$$

Therefore, Z_2 is a weighted modified renewal process (see Cox [Co62]).

The characteristic function $G_2(\omega)$ of the r.v. Z_2 is defined as

$$G_2(\omega) \triangleq E[e^{j\omega Z_2}] = \int e^{j\omega z_2} dH_2(z_2) \quad (3.4)$$

Using Eq. 3.3, we find

$$G_2(\omega) = (1 - \alpha) \sum_{n=0}^{\infty} \alpha^n \int e^{j\omega z_2} d(\{F_1(z_2)\}^n \otimes F_2(z_2)) \\ = (1 - \alpha) \sum_{n=0}^{\infty} \alpha^n \phi_1^n(\omega) \phi_2(\omega) \quad (3.5)$$

where $\phi_i(\omega)$ is the characteristic function of the r.v. X_i ($i=1,2$).

The last equation is a result of the independency of X_1 and X_2 , and the convolution properties (see Papoulis [Pa65]).

Therefore,

$$G_2(\omega) = \frac{(1-\alpha)\phi_2(\omega)}{1 - \alpha\phi_1(\omega)} \quad (3.6)$$

Since $\phi_1(\omega)$ and $\phi_2(\omega) \rightarrow 1$ as $\omega \rightarrow 0$,

then $G_2(\omega) \rightarrow 1$ as $\omega \rightarrow 0$.

Having obtained the characteristic function, we can obtain the n^{th} moment as follows

$$\overline{Z_2^n} = \frac{1}{i^n} \left. \frac{d^n G_2(\omega)}{d\omega^n} \right|_{\omega=0} \quad (3.7)$$

Of interest to us are the first and second moments.

The first moment $\overline{Z_2}$ represents the mean interdeparture time and is given as

$$\begin{aligned} \overline{Z_2} &= \frac{1}{i} \left. \frac{d G_2(\omega)}{d\omega} \right|_{\omega=0} \\ &= \frac{1}{i} \left. \frac{(1-\alpha)\phi_2'(\omega)[1-\alpha\phi_1(\omega)] + \alpha(1-\alpha)\phi_2(\omega)\phi_1'(\omega)}{[1-\alpha\phi_1(\omega)]^2} \right|_{\omega=0} \end{aligned} \quad (3.8)$$

$$\phi_1'(0) = \left. \frac{d\phi_1(\omega)}{d\omega} \right|_{\omega=0} = i\mu_1^{-1} \quad (3.9)$$

$$\phi_2'(0) = \left. \frac{d\phi_2(\omega)}{d\omega} \right|_{\omega=0} = i\mu_2^{-1} \quad (3.10)$$

$$\text{and } \phi_1(0) = \phi_2(0) = 1$$

Therefore,

$$\mu_{d_2}^{-1} \triangleq \overline{Z_2} = \mu_2^{-1} + \frac{\alpha}{1-\alpha} \mu_1^{-1} \quad (3.11)$$

From this expression, we see that the departure rate of class 2 (μ_{d_2}) is equal to the service rate of class 2 (μ_2) if $\alpha = 0$ (only class 2 exists in the system), and $\mu_{d_2} = 0$ if $\alpha = 1$ (no class 2 in the system). When both

classes exist in the system ($0 < \alpha < 1$), the departure rate μ_{d_2} will take values in the range $[0, \mu_2]$ according to Eq. 3.11.

Similarly for the second moment $\overline{Z_2^2}$

$$\overline{Z_2^2} = \frac{1}{i^2} \left. \frac{d^2 G_2(\omega)}{d\omega^2} \right|_{\omega=0} \quad (3.12)$$

Performing the derivation and noting that

$$\phi_1'(0) = \left. \frac{d \phi_1^2(\omega)}{d\omega^2} \right|_{\omega=0} = i^2 \overline{X_1^2} = i^2 (\sigma_1^2 + (\mu_1^{-1})^2) \quad (3.13)$$

$$\phi_2''(0) = \left. \frac{d \phi_2^2(\omega)}{d\omega^2} \right|_{\omega=0} = i^2 \overline{X_2^2} = i^2 (\sigma_2^2 + (\mu_2^{-1})^2) \quad (3.14)$$

We find

$$\begin{aligned} \overline{Z_2^2} = \frac{1}{(1-\alpha)^2} \{ & (1-\alpha)^2 \overline{X_2^2} + \alpha(1-\alpha) \overline{X_1^2} + 2 \alpha(1-\alpha) \mu_1^{-1} \mu_2^{-1} \\ & + 2 \alpha^2 (\mu_1^{-1})^2 \} \end{aligned} \quad (3.15)$$

The variance of Z_2 is

$$\begin{aligned} \sigma_{d_2}^2 \triangleq \sigma_{Z_2}^2 &= \overline{Z_2^2} - (\overline{Z_2})^2 \\ &= \frac{1}{(1-\alpha)^2} \{ (1-\alpha)^2 \sigma_2^2 + \alpha(1-\alpha) \sigma_1^2 + \alpha (\mu_1^{-1})^2 \} \end{aligned} \quad (3.16)$$

and the squared coefficient of variation defined as

$$c_{d2} = \frac{\sigma_{z2}^2}{(\bar{z}_2)^2} = \frac{(1-\alpha)^2 \frac{c_2}{\beta^2} + ((1-\alpha)c_1 + 1)\alpha}{\left(\alpha + \frac{(1-\alpha)}{\beta}\right)^2} \quad (3.17)$$

It should be noted that if the service times for the two classes are exponentially distributed and have the same mean (i.e. $c_1 = c_2 = 1$ and $\beta = 1$), then $c_{d2} = 1$. This means that the interdeparture times of the low priority class are exponentially distributed, and the departure process is Poisson. However, if $c_1 = c_2 = 1$ but $\beta \neq 1$, the departure process is not Poisson any more, but is a renewal process.

Though Eqs. 3.6, 3.11 and 3.17 are calculated for Poisson arrivals, they still are applicable in the limit if arrivals are independent renewal processes. For this case, the number of arrivals of class 1 and 2 in time t approaches asymptotically the normal distribution with means $\lambda_1 t$ and $\lambda_2 t$ respectively [Co62]. Therefore, the probability of an arrival of class 1 approaches $\lambda_1 t / (\lambda_1 + \lambda_2)t = \alpha$, and that of class 2 approaches $1 - \alpha$ as $t \rightarrow \infty$. Since the departure sequence is equal to the arrival sequence, the results of these sections are applicable to the GI/G/1 queueing systems in the limit as $t \rightarrow \infty$ (i.e. in the steady state).

3.3 The Diffusion Equation and the Steady State Solutions

In this section, we derive expressions for the diffusion equations and the steady state solutions for the Preemptive-Resume priority service discipline.

The Diffusion Equation

For the high priority class under heavy traffic conditions, the arrival and departure processes are renewal with means and squared coefficients of variation given as $(\lambda_1^{-1}, c_{a_1})$ and (μ_1^{-1}, c_1) , respectively. Note that the interdeparture times here are the same as the service times. Using Eqs. 2.26 and 2.27, the change in the number of customers of class 1 in the system between the time t and $t + dt$ is approximately normally distributed with mean and variance given as

$$E[\Delta Q_1(t)] = (\lambda_1 - \mu_1)dt = b_1 dt \quad (3.18)$$

$$\text{var}[\Delta Q_1(t)] = (c_{a_1} \lambda_1 + c_1 \mu_1)dt = a_{11} dt \quad (3.19)$$

For the low priority class let

$A_2(t)$ represent the cumulative number of arrivals of class 2 customers up to time t ,

$D_2(t)$ represent the cumulative number of departures of class 2 customers up to time t ,

and $Q_2(t)$ represent the number in the system of class 2 customers at time t (including the one in service, if any).

$$\text{Therefore, } Q_2(t) \triangleq A_2(t) - D_2(t) \quad (3.20)$$

The change in the number of customers of class 2 in the system between time t and $t + dt$ is

$$\begin{aligned}\Delta Q_2(t) &\stackrel{\Delta}{=} Q_2(t + dt) - Q_2(t) \\ &= \Delta A_2(t) - \Delta D_2(t)\end{aligned}\quad (3.21)$$

where

$$\begin{aligned}\Delta A_2(t) &\stackrel{\Delta}{=} A_2(t + dt) - A_2(t) \\ \Delta D_2(t) &\stackrel{\Delta}{=} D_2(t + dt) - D_2(t)\end{aligned}$$

The interarrival times and the interdeparture times are both i.i.d. with the (mean, variance) given by $(\lambda_2^{-1}, \sigma_{a_2}^2)$ and $(\mu_{d_2}^{-1}, \sigma_{d_2}^2)$, respectively. If we assume dt is sufficiently large for many arrivals and departures to take place, also that $Q_2(t)$ does not go to zero in this interval, then from the central limit theorem $\Delta Q_2(t)$ is approximately normally distributed with the following mean and variance

$$E[\Delta Q_2(t)] = (\lambda_2 - \mu_{d_2})dt = b_2 dt \quad (3.22)$$

$$\text{var}[\Delta Q_2(t)] = (c_{a_2} \lambda_2 + c_{d_2} \mu_{d_2})dt = a_{22} dt. \quad (3.23)$$

Now, we know that the distribution of the changes in the number of customers in the system of the high priority class in time dt is completely independent of the low priority class, as we explained before, therefore

$$\text{cov}[\Delta Q_1(t), \Delta Q_2(t)] = \text{cov}[\Delta Q_2(t), \Delta Q_1(t)] = 0 \quad (3.24)$$

$$\text{i.e. } a_{12} = a_{21} = 0$$

Given the initial state x_0 at $t = 0$, then the conditional probability density at time t , $p(\bar{x}, t)$ satisfies the 2-dimensional diffusion equation

$$\frac{\partial}{\partial t} p(\bar{x}, t) = \frac{1}{2} \sum_{i=1}^2 a_{ii} \frac{\partial^2}{\partial x_i^2} p(\bar{x}, t) - \sum_{i=1}^2 b_i \frac{\partial}{\partial x_i} p(\bar{x}, t) \quad (3.25)$$

where $\bar{X}(t)$ is the 2-dimensional continuous process that approximates the discrete vector process $Q(t)$

Boundary Conditions and Steady State Solutions

Since $X(t)$ is a positive-valued vector, we introduce 2 reflecting boundaries at $x_1 = 0$ and $x_2 = 0$. This is done by letting the directional fluxes ψ_1 , and ψ_2 equal to zero, i.e.,

$$-\frac{1}{2} a_{11} \frac{\partial p(\bar{x}, t)}{\partial x_1} + b_1 p(\bar{x}, t) = 0 \Big|_{x_1=0} \quad (3.26)$$

$$-\frac{1}{2} a_{22} \frac{\partial p(\bar{x}, t)}{\partial x_2} + b_2 p(\bar{x}, t) = 0 \Big|_{x_2=0} \quad (3.27)$$

The steady state solution can be obtained by setting $\frac{\partial p(\bar{x}, t)}{\partial t}$ equal to zero in the diffusion equation, and getting a solution that satisfies the differential equation and the boundary conditions, i.e. by solving the equation

$$\frac{\partial \psi_1}{\partial x_1} = - \frac{\partial \psi_2}{\partial x_2} \quad (3.28)$$

subject to $\psi_1 \Big|_{x_1=0} = \psi_2 \Big|_{x_2=0} = 0$

where $\psi_1 \triangleq \psi_1(x_1 | x_2; t)$

and $\psi_2 \triangleq \psi_2(x_2 | x_1; t)$

$$\text{Let } \frac{\partial \psi_1}{\partial x_1} = - \frac{\partial \psi_2}{\partial x_2} = k \text{ (constant)} \quad (3.29)$$

$$\begin{aligned} \text{Then by simple integration; } \psi_1 &= kx_1 + A \\ \text{and } \psi_2 &= -kx_2 + B \end{aligned} \quad (3.30)$$

To satisfy the boundary conditions A and B must be equal to zero.

Furthermore, to have a finite flux as x_1 or $x_2 \rightarrow \infty$, then k must also be equal to zero.

The problem now is equivalent to solving the following two simultaneous partial differential equations of first order

$$\frac{a_{11}}{2} \frac{\partial p(\bar{x})}{\partial x_1} - b_1 p(\bar{x}) = 0 \quad (3.31)$$

$$\frac{a_{22}}{2} \frac{\partial p(\bar{x})}{\partial x_2} - b_2 p(\bar{x}) = 0$$

$$\text{where } p(\bar{x}) = p(\bar{x}, \infty)$$

It is clear in this case, we can try the solution $p(\bar{x}) = p(x_1) p(x_2)$, which will lead to two separate differential equations of first order

$$\frac{a_{11}}{2} \frac{dp(x_1)}{dx_1} - b_1 p(x_1) = 0 \quad (3.32)$$

$$\frac{a_{22}}{2} \frac{dp(x_2)}{dx_2} - b_2 p(x_2) = 0 \quad (3.33)$$

$$p(x_1), p(x_2) \neq 0$$

Therefore, $p(x_1)$ is given by Eq. 2.30, and the approximate distribution of the number of customers in the system of class 1 is obtained by discretizing $p(x_1)$ to get

$$\hat{p}(n_1) = (1 - \hat{\rho}_1) \hat{\rho}_1^{n_1} \quad (3.34)$$

where

$$\hat{\rho}_1 = \exp \frac{-2(1 - \rho_1)}{c_{a1} \rho_1 + c_1} \quad (3.35)$$

$$\hat{\rho}_2 = \exp \frac{-2(1 - \rho_{2\text{eff}})}{c_{a2} \rho_{2\text{eff}} + c_{d2}} \quad (3.36)$$

and

$$\rho_1 = \lambda_1 / \mu_1 \quad (3.37)$$

$$\begin{aligned} \rho_{2\text{eff}} &= \frac{\lambda_2}{\mu_{d2}} = \frac{\lambda_2}{1-\alpha} [(1-\alpha)\mu_2^{-1} + \alpha\mu_1^{-1}] \\ &= \frac{\lambda_2}{\mu_2} + \frac{\lambda_1}{\mu_1} \\ &= \rho_2 + \rho_1 \end{aligned} \quad (3.38)$$

For class 1 customers, we improve the solution by making

$$\hat{p}(n_1) = \begin{cases} 1 - \rho_1 & \text{for } n_1 = 0 \\ \rho_1 (1 - \hat{\rho}_1) \hat{\rho}_1^{n_1-1} & n_1 \geq 1 \end{cases} \quad (3.39)$$

3.4 Analysis of Results

3.4.1 Introduction

In this section, we test the accuracy of the diffusion approximations for Preemptive-Resume queueing systems. We first check the accuracy of the mean number in an M/G/1 queueing system using existing analytical formulas. Following this, we proceed to check the accuracy of the diffusion approximations for the general GI/G/1 queueing system by comparing the mean, the variance and the distribution of the number in the system with GPSS simulation results. We restrict the following discussions to the low priority class since we already know that the accuracy of the diffusion approximation of the high priority class is similar to that of GI/G/1 with one class which was examined by Reiser and Kobayashi [RK74].

3.4.2 Accuracy of the Diffusion Approximation for M/G/1 Queueing Systems

The mean number of the low priority customers in an M/G/1 queueing system with Preemptive-Resume discipline is derived in many references [e.g. Ja68] and is given by

$$E[n_2] = \frac{\lambda_2}{1 - \rho_1} [\mu_2^{-1} + \frac{\lambda_1 E[S_1^2] + \lambda_2 E[S_2^2]}{2(1 - \rho_1 - \rho_2)}] \quad (3.40)$$

where ρ_1 is the utilization of the server by class 1

$E[S_1^2]$ is the second moment of the service time of class 1

and $\rho_1 + \rho_2 < 1$, i.e. system is stationary

We can write Eq. 3.40 as

$$E[n_2] = \frac{1}{1 - \rho_1} \left[\rho_2 + \frac{\rho_1 \rho_2 \beta (1 + c_1) + \rho_2^2 (1 + c_2)}{2(1 - \rho_1 - \rho_2)} \right] \quad (3.41)$$

where

$$\beta = \frac{\mu_2}{\mu_1} = \frac{(1 - \alpha)\rho_1}{\alpha\rho_2} \quad (3.42)$$

The mean and the variance of the number in the system of the low priority class are calculated from the diffusion approximation as

$$E[\hat{n}_2] = \hat{\rho}_2 / (1 - \hat{\rho}_2) \quad (3.43)$$

$$\text{var}[\hat{n}_2] = \hat{\rho}_2 / (1 - \hat{\rho}_2)^2 \quad (3.44)$$

$$\text{where} \quad \hat{\rho}_2 = \exp \left\{ \frac{-2(1 - \rho_{2\text{eff}})}{\rho_{2\text{eff}} c_{a2} + c_{d2}} \right\} \quad (3.45)$$

and

$$\rho_{2\text{eff}} = \rho_1 + \rho_2$$

The relative error in the mean $E[\hat{n}_2]$ and the variance $\text{var}[\hat{n}_2]$ are defined as

$$\text{ER}_2 = \frac{E[n_2] - E[\hat{n}_2]}{E[n_2]} \times 100 \quad (3.46)$$

$$\text{ERV}_2 = \frac{\text{var}[n_2] - \text{var}[\hat{n}_2]}{\text{var}[n_2]} \times 100 \quad (3.47)$$

Tables 3.1 and 3.2 show the effect of changing c_2 on the accuracy for fixed values of α , β , c_1 and ρ_{eff} . We have the following observations:

- The relative error is in general small (less than 6%) and the diffusion approximation is "satisfactory" for all the cases studied.
- We tend in general to underestimate if service distribution is Erlang ($c_2 < 1$) and overestimate if it is hyperexponential ($c_2 > 1$). These trends are similar to those observed for the single class M/G/1 queueing systems [RK74].
- The relative error increases as we deviate from the exponential service distribution (i.e. deviate from $c_2 = 1$).

Table 3.3 shows the effect of changing the total utilization, while keeping the utilization of the high priority class constant. It should be noted that the distributions of the service times are assumed exponential and α is constant.

Table 3.4 shows the effect of changing α while keeping the utilization of both classes fixed and assuming that the service time distributions for both classes are exponential.

From the two tables, we find that:

- The relative error decreases as the total utilization ρ_{eff} increases to approach the unity.
- The relative error decreases with the increase of the value α or the decrease of β .

c_2	$E[n_2]$	$E[\hat{n}_2]$	ER_2
.5	6.33	6.17	2.5
1	8.11	8.15	-.4
2	11.7	12.1	-3.6
3	15.2	16	-5.4

Table 3.1 Effect of Changing c_2 on the Accuracyfor $\alpha = .5$, $\beta = .125$, $c_1 = 1$ and $\rho_{\text{eff}} = .9$

c_2	$E[n_2]$	$E[\hat{n}_2]$	ER_2
.5	6.48	6.34	2.22
1	8.26	8.31	-.61
2	11.8	12.3	-3.75
3	15.4	16.2	-5.44

Table 3.2 Effect of Changing c_2 on the Accuracyfor $\alpha = .3$, $\beta = .292$, $c_1 = 1$ and $\rho_{\text{eff}} = .9$

ρ_{eff}	β	$E[n_2]$	$E[\hat{n}_2]$	ER_2
.9	.667	8.17	8.43	-3.3
.95	.622	17.3	17.7	-2.4
.98	.598	44.8	45.7	-2.
.999	.583	9160	9330	-1.8

Table 3.3 Effect of Changing ρ_{eff} (or β) on the Accuracy
for $\alpha = .3$, $\rho_1 = .2$ and $c_1 = c_2 = 1$

α	β	$E[n_2]$	$E[\hat{n}_2]$	ER_2
.1	1.125	9.	9.13	-1.47
.3	.292	8.26	8.31	-.61
.5	.125	8.11	8.15	-.42
.7	.054	8.05	8.08	-.32
.9	.014	8.01	8.04	-.29

Table 3.4 Effect of Changing α or β on the Accuracy
for $\rho_1 = .1$, $\rho_2 = .8$, $c_1 = 1$ and $c_2 = 1$

3.4.3 Accuracy of the Diffusion Approximation for GI/G/1 Queueing Systems

We now compare the results obtained from the diffusion approximation with those obtained from simulation runs. The objective is to support and complement the results reported in sec. 3.4.2. We briefly describe the technique used to obtain confidence interval for the mean in the simulation runs. A model for an h-stage service facility is then described, and the first and the second moments are obtained. Results are then presented for four cases, the first is an M/M/1 system, the second is an M/G/1 system, and the last two are GI/G/1 queueing systems. The use of simulation in this section allowed us not only to evaluate the accuracy of the mean as in the previous section, but also to compare the distribution of the number of customers in the system.

Confidence Interval

A confidence interval, at a level of confidence of $(1 - d)$, is defined as the interval within which the result of a simulation run will lie with a probability $(1 - d)\%$. In other words, there is a probability of $d\%$ that a simulation result will be outside the confidence interval.

There are several techniques for the evaluation of the variance of the estimated mean, and therefore the confidence intervals [Kj75]. Examples of these techniques are replication of simulation runs, dividing a run to independent batches or subruns, autocorrelation techniques, and the technique of independent cycles or tours. The last technique (independent cycles) was first presented by Kabak in 1968 [Ka68], who pointed out that in a single server queueing system the arrival of a customer to an empty system starts a cycle which is independent from the past. The state of the system that corresponds to the start of a cycle

is called a regeneration point and may be any state which is visited frequently. These independent cycles are identically distributed and are used to calculate the confidence interval. Crane and Iglehart [CI74] applied the technique of independent cycles to the case of multi-server system. In this case, cycles are assumed to start with the arrival of a customer to an empty system (all servers are idle). The technique is used to estimate confidence intervals for a number of parameters.

For illustrative purposes, we consider the mean waiting time $E[w]$ determined by

$$E[w] = \frac{\bar{Y}}{\bar{N}} = \frac{\sum_{i=1}^M Y_i}{\sum_{i=1}^M m_i} \quad (3.48)$$

where

$Y_i = \sum_{k=1}^{m_i} w_{ik}$ is the sum of waiting time of all customers in cycle i

w_{ik} is the waiting time of customer k in cycle i

M is the number of cycles

m_i is the number of customers in cycle i

Assuming large value for M , Crane and Iglehart [CI74] determine the $(1 - d)\%$ confidence interval for the mean waiting time as

$$\left[\frac{(\bar{Y} \bar{N} - Ks_{12}) - D^{\frac{1}{2}}}{(\bar{N})^2 - Ks_{22}}, \frac{(\bar{Y} \bar{N} - Ks_{12}) + D^{\frac{1}{2}}}{(\bar{N})^2 - Ks_{22}} \right] \quad (3.49)$$

where

$$s_{11} = \text{var } Y_i$$

$$s_{12} = \text{cov } (Y_i, m_i)$$

$$s_{22} = \text{var } (m_i)$$

$K = (Z^{d/2})^2/M$, and $Z^{d/2}$ is the upper $d/2$ point of standard (0,1) normal

$$D = (\bar{Y} \bar{N} - Ks_{12})^2 - (\bar{Y}^2 - Ks_{11})(\bar{N}^2 - Ks_{22})$$

Using the previous confidence interval for $E[W]$, we can obtain an approximate confidence interval for

$$E[Q] = \lambda E[W] + \rho \quad (3.50)$$

For more details we refer the reader to Crane and Iglehart [CI74].

Service Facility with h-stages

The service facility is composed of h -stages in tandem as shown in fig. 3.2. Stage " i " serving customers at a rate μ_i and the service time has a negative exponential distribution. However, customers demand service from the i^{th} stage with probability π_i , and skip the stage with probability $(1 - \pi_i)$. The reason for choosing this model for the service facility is to permit different function partitioning between the two processors of the Host processor and the Front-End Processor interface configuration discussed in Chapters 1 and 7.

Fig. 3.2 h -Stage Service Facility

We proceed now to calculate the distribution of the service time and the corresponding first and second moments.

The Laplace transform for the service time pdf is

$$\begin{aligned} T(s) &= \prod_{i=1}^h \left\{ \pi_i \frac{\mu_i}{s + \mu_i} + (1 - \pi_i) \right\} \\ &= \prod_{i=1}^h \frac{s(1 - \pi_i) + \mu_i}{s + \mu_i} \end{aligned} \quad (3.51)$$

The n^{th} moment of $T(s)$ is calculated as [K175]

$$E[T^n] = \frac{1}{(j)^n} \left. \frac{d^n T(s)}{ds^n} \right|_{s=0} \quad (3.52)$$

Therefore, the mean is

$$\begin{aligned} \bar{T} &= - \left. \frac{d}{ds} T(s) \right|_{s=0} \\ &= \sum_{i=1}^h \frac{\pi_i}{\mu_i} \end{aligned} \quad (3.53)$$

and the second moment is given as

$$\begin{aligned} E[T^2] &= \left. \frac{d^2}{ds^2} T(s) \right|_{s=0} \\ &= \sum_{i=1}^h \frac{2\pi_i}{\mu_i^2} + \sum_{k=1}^h \sum_{i=1, i \neq k}^h \frac{\pi_i \pi_k}{\mu_i \mu_k} \end{aligned} \quad (3.54)$$

Simulation

The simulation program is written in GPSS language. The program consists of approximately 40 blocks, and was run in all cases for 25,000 transactions (a GPSS transaction represents a customer in the queueing system). The CPU time required to run the simulation on the Carleton University XDS SIGMA-9 computer is estimated to be 3.3 minutes on the average. The simulation program stores and keeps track of all the details that are necessary to obtain confidence interval for the mean number in the system. The flowchart of the GPSS simulation program is presented in Appendix A.

Numerical Results

Table 3.5 and figure 3.3 present the results of an M/M/1 queueing system. The value of α is 0.5, which means that the mean arrival rate of both classes is the same. As will be shown later, such a value of α is applicable in modelling Front-End Processors. The ratio of service rates β is equal to .125 which means that mean service time of the low priority class is 8 times that of the high priority class. The total utilization is 0.9.

In case 2, we change the service facility of the low-priority class to the 2-stage (H_2) facility with $C_2 = 1.564$. Table 3.6 and figure 3.4 summarize the results of this case.

In these two cases it is interesting to notice that the diffusion approximation estimates are more accurate than the simulation results with respect to the exact analytical results. We can obtain more accurate simulation results by running the simulation for a longer time; however, the computer time cost is prohibitive.

Poisson arrival

 $\alpha = .5$ Exponential service times $(\mu_1^{-1} = 4, \mu_2^{-1} = 32)$ $\beta = .125$ $(c_1 = c_2 = 1)$ $\rho_1 = .1 \quad \rho_2 = .8 \quad c_{d2} = .83 \quad \hat{\rho}_2 = .888$

	Simulation	Diffusion	Analytical
mean	7.49	7.97	7.93
90% conf. int.	[6.29 - 8.42]		
for mean			
var	56.13	71.16	
ERV ₂		-26.8	

Table 3.5 Accuracy of the Diffusion Approximation for Case 1

Poisson arrival $\alpha = .5$
 Exponential, and 2-stage (H_2) ($\mu_1^{-1} = 4, \mu_2^{-1} = 32.8$) $\beta = .122$
 service times ($c_1 = 1, c_2 = 1.564$)

$\rho_1 = .1$ $\rho_2 = .82$ $c_{d_2} = 1.278$ $\hat{\rho}_2 = .93$

	Simulation	Diffusion	Analytical
mean	14.53	13.2	13
90% conf. int.	[12.2 - 16.33]		
for mean			
var	347.45	188.57	
ERV ₂		45.7	

Table 3.6 Accuracy of the Diffusion Approximation for Case 2

Fig. 3.3 Distribution of the Number in the System of the Low Priority Customers for Case 1

Fig. 3.4 Distribution of the Number in the System of the Low Priority Customers for Case 2

The arrival process of class 2 is changed in the case reported in table 3.7 and fig. 3.5 to a uniform distribution with interarrival time in the range 30 to 50 inclusive.

Finally we present another case with 2 stage (H_2) service facility and uniform distributed interarrival time [30-70] and different values for α and β . Results are summarized in table 3.8 and fig. 3.6.

In all these cases, we reported the values of the mean number in the system calculated from Eq. 3.41 for M/G/1 queueing system which was used in sec. 3.4.2. It can be observed from cases 3 and 4 that the M/G/1 expression is not satisfactory for GI/G/1, i.e. the indiscriminate assumption of Poisson arrivals in GI/G/1 queues will lead to serious errors.

The above analysis indicates that results obtained by the diffusion approximation for GI/G/1 Preemptive-Resume queueing systems are generally satisfactory as they contain relatively small errors.

3.5 Concluding Remarks

In this chapter, we studied the diffusion approximation of a Preemptive-Resume GI/G/1 queueing system with two classes of customers. We derived approximate expressions for the distribution of the inter-departure times and the number in the system of the low priority customers. The accuracy of the approximation was tested against analytical formulas and/or simulation and proved to be satisfactory in all the cases considered.

Poisson, and uniform arrivals $c_{a1} = 1, c_{a2} = 1/48$ $\alpha = .5$

Exponential, and 2-stage (H_2) service times $(\mu_1^{-1} = 4, \mu_2^{-1} = 32.8)$ $\beta = .122$

$(c_1 = 1, c_2 = 1.564)$

$\rho_1 = .1 \quad \rho_2 = .803 \quad c_{d2} = 1.274 \quad \hat{\rho}_2 = .861$

	Simulation	Diffusion	Another approximation (Analytical M/G/1)
mean	6.43	6.17	10.5
ER ₂		4	-63.3
90% conf. int. for mean	[4.97 - 7.61]		
var	45.83	44.07	
ERV ₂		3.8	

Table 3.7 Accuracy of the Diffusion Approximation for Case 3

Poisson, and uniform arrivals $c_{a_1} = 1, c_{a_2} = 4/75$ $\alpha = .33$

Exponential, and 2-stage (H_2) service times $(\mu_1^{-1} = 25, \mu_2^{-1} = 32.8)$ $\beta = .8$

$(c_1 = 1, c_2 = 1.564)$

$\rho_1 = 0.25$ $\rho_2 = .64$ $c_{d_2} = 1.208$ $\hat{\rho} = .839$

	Simulation	Diffusion	Another approximation (Analytical M/G/1)
mean	5.52	5.22	8.76
ER ₂		5.4	-58.7
90% conf. int.	[4.82 - 6.12]		
for mean			
var	32.3	32.6	
ERV ₂		- .9	

Table 3.8 Accuracy of the Diffusion Approximation for Case 4

Fig. 3.5 Distribution of the Number in the System of the Low Priority Customers for Case 3

Fig. 3.6 Distribution of the Number in the System of the Low Priority Customers for Case 4

4. DIFFUSION APPROXIMATION FOR AN FCFS SINGLE SERVER WITH TWO CLASSES

4.1 Introduction

In this chapter we derive the diffusion equation for GI/G/1 queueing systems with two classes of customers served with different rates but on a first-come-first-served (FCFS) basis. In sec. 4.2, we obtain expressions for the distribution of the interdeparture times of each class for an M/G/1 system, and we show that these results are applicable, in the limit, to a GI/G/1 system. The diffusion equation is then derived in sec. 4.3, and its solution with reflecting boundaries is obtained in sec. 4.4. Analysis of the results is presented in sec. 4.5.

4.2 The Departure Process

Consider a single server that receives two classes of customers, and assume arrivals are Poisson with rates λ_1 and λ_2 , respectively. The server gives no priority to any particular class, but services the customers on FCFS basis. However, the service time distributions are different for each class. Assume class 1 customers are served at a rate μ_1 , and variance for service times of σ_1^2 . Similarly for class 2, the service rate is μ_2 , and the variance for service times is σ_2^2 . Let $F_1(x_1)$ represent the cumulative distribution function (PDF) of the service times of class 1 customers, and $\phi_1(\omega)$ be the corresponding characteristic function defined as

$$\phi_1(\omega) \triangleq E[e^{j\omega X_1}] = \int_{-\infty}^{\infty} e^{j\omega x_1} dF_1(x_1) \quad (4.1)$$

The corresponding functions for class 2 customers are $F_2(x_2)$ and $\phi_2(\omega)$.

Assume Z_1 is the r.v. that represents the interdeparture times of class 1, $H_1(z_1)$ is the PDF of Z_1 , and $G_1(\omega)$ is the characteristic function of Z_1 .

Since we assumed the arrivals of the two classes are Poisson and independent, then the total arrival to the queue is likewise Poisson with a mean rate $\lambda = \lambda_1 + \lambda_2$. In addition, the probability of an arrival of class 1 being observed is α , and the probability of an arrival of class 2 being observed is $(1 - \alpha)$, where $\alpha = \lambda_1/\lambda$. Furthermore, the service is on an FCFS basis, therefore the departure sequence is equal to the arrival sequence. We also assume that the queueing system experiences heavy traffic (i.e. the queue is never empty).

Let us now observe the queueing system immediately after a departure of a class 1 customer has taken place and wait for the next departure of a class 1 customer. This may be the first departure to be observed from the queueing system, or may follow one or more departures of class 2 customers.

In fact there may be n departures of class 2, $0 \leq n \leq \infty$, contained between two successive departures of class 1 customers. These sequences of departures are mutually exclusive events. Thus, we can use the law of total probability to calculate the PDF of the r.v. Z_1 as follows:

$$\begin{aligned} P[Z_1 \leq z_1] &= P[Z_1 \leq z_1 \text{ and } \{1\}] \\ &\quad + P[Z_1 \leq z_1 \text{ and } \{2,1\}] \\ &\quad + P[Z_1 \leq z_1 \text{ and } \{2,2,1\}] + \dots \end{aligned} \quad (4.2)$$

where $\{1\}$ is the sequence of departures when the first one is a class 1 customer,

$\{2,1\}$ is the sequence of departures consisting of a class 2 customer followed by a class 1 customer.

etc.

$$\begin{aligned} P[Z_1 \leq z_1] &= P[X_1 \leq z_1 \text{ and } \{1\}] + P[X_1 + X_2^1 \leq z_1 \text{ and } \{2,1\}] \\ &+ P[X_1 + X_2^1 + X_2^2 \leq z_1 \text{ and } \{2,2,1\}] + \dots \end{aligned}$$

where X_1 is the r.v. with PDF $F_1(x_1)$

and $\{X_2^1, X_2^2, X_2^3, \dots\}$ are independent non-negative r.v.'s all with the same PDF $F_2(x_2)$.

Since the sequence of departures is equal to the sequence of arrivals, then

$$P\{1\} = \alpha$$

$$P\{2,1\} = (1 - \alpha)\alpha$$

$$P\{2,2,1\} = (1 - \alpha)^2\alpha \quad (4.3)$$

etc.

Also we know that

$$P[X_1 \leq z_1] = F_1(z_1)$$

$$P[X_1 + X_2^1 \leq z_1] = F_1(z_1) \otimes F_2(z_1)$$

$$\begin{aligned} P[X_1 + X_2^1 + X_2^2 \leq z_1] &= F_1(z_1) \otimes F_2(z_2) \otimes F_2(z_2) \\ &= F_1(z_1) \otimes \{F_2(z_2)\}^2 \end{aligned} \quad (4.4)$$

etc.

Furthermore, the departure sequences are independent of the service times X_1 and X_2 , therefore

$$\begin{aligned}
 H_1(z_1) &\triangleq P[Z_1 \leq z_1] = \alpha F_1(z_1) + \alpha(1-\alpha)F_1(z_1) \otimes F_2(z_1) \\
 &\quad + \alpha(1-\alpha)^2 F_1(z_1) \otimes \{F_2(z_1)\}^2 + \dots \\
 &= \alpha \sum_{n=0}^{\infty} (1-\alpha)^n F_1(z_1) \otimes \{F_2(z_1)\}^n
 \end{aligned} \tag{4.5}$$

which implies that the r.v. Z_1 is a weighted modified renewal process. It is easy to show that the characteristic function $G_1(\omega)$ of the r.v. Z_1 is given by

$$G_1(\omega) = \frac{\alpha \phi_1(\omega)}{1 - (1-\alpha)\phi_2(\omega)} \tag{4.6}$$

We can obtain the n^{th} moment in the following way:

$$\overline{Z_1^n} = \frac{1}{i^n} \left. \frac{d^n G_1(\omega)}{d\omega^n} \right|_{\omega=0} \tag{4.7}$$

Therefore, the first and second moments are given by

$$\mu_{d_1}^{-1} \triangleq \overline{Z_1} = \mu_1^{-1} + \frac{1-\alpha}{\alpha} \mu_2^{-1} \tag{4.8}$$

$$\begin{aligned}
 \overline{Z_1^2} &= \frac{1}{\alpha^2} \{ \alpha^2 \overline{X_1^2} + \alpha(1-\alpha) \overline{X_2^2} + \\
 &\quad 2\alpha(1-\alpha) \mu_1^{-1} \mu_2^{-1} + 2(1-\alpha)^2 (\mu_2^{-1})^2 \}
 \end{aligned} \tag{4.9}$$

The variance of the r.v. Z_1 is

$$\begin{aligned}\sigma_{d_1}^2 &\triangleq \sigma_{Z_1}^2 = \overline{Z_1^2} - (\overline{Z_1})^2 \\ &= \frac{\mu}{\alpha^2} \{ \alpha^2 \sigma_1^2 + \alpha(1-\alpha)\sigma_2^2 + (1-\alpha)(\mu_2^{-1})^2 \}\end{aligned}\quad (4.10)$$

and the squared coefficient of variation of the interdeparture times Z_1 is given by

$$c_{d_1} \triangleq \frac{\sigma_{Z_1}^2}{(\overline{Z_1})^2} = \frac{\alpha^2 c_1 + (\alpha c_2 + 1) \frac{(1-\alpha)}{\beta^2}}{(\alpha + \frac{(1-\alpha)}{\beta})^2} \quad (4.11)$$

where c_1 is the squared coefficient of variation of the r.v. X_1 and is equal to $\sigma_1^2 \mu_1^2$,

$$c_2 \triangleq \sigma_2^2 \mu_2^2,$$

and $\beta \triangleq \frac{\mu_2}{\mu_1}$ is the ratio of the mean service rates of class 2 to class 1.

The corresponding equations for the mean, the variance, and the squared coefficient of variation of the interdeparture times of class 2 customers are similar to those of the interdeparture times of the low priority class of chapter 3. They are given by Eqs. 3.11, 3.16 and 3.17 as:

$$\mu_{d_2}^{-1} = \overline{Z_2} = \mu_2^{-1} + \frac{\alpha}{1-\alpha} \mu_1^{-1} \quad (4.12)$$

$$\sigma_{d_2}^2 = \sigma_{Z_2}^2 = \frac{1}{(1-\alpha)^2} \{ (1-\alpha)^2 \sigma_2^2 + \alpha(1-\alpha)\sigma_1^2 + \alpha(\mu_1^{-1})^2 \} \quad (4.13)$$

$$c_{d_2} = \frac{(1-\alpha)^2 \frac{c_2}{\beta^2} + ((1-\alpha)c_1 + 1)\alpha}{(\alpha + \frac{1-\alpha}{\beta})^2} \quad (4.14)$$

Notice that $\mu_{d_1}/\mu_{d_2} = \lambda_1/\lambda_2 = \alpha/(1-\alpha)$, i.e. the ratio of the mean departure rates of the two classes is equal to the ratio of their mean arrival rates. In addition, note that if the service times for the two classes are exponential and have the same mean (i.e. $c_1 = c_2 = 1$ and $\beta = 1$), then $c_{d_1} = c_{d_2} = 1$, which implies that the departure processes are Poisson. However, if $c_1 = c_2 = 1$ but $\beta \neq 1$, the departure processes are not Poisson any more, but are renewal processes.

As we explained in the previous chapter, all these results apply to the GI/G/1 queuing systems in the limit as $t \rightarrow \infty$ (i.e. in the steady state).

4.3 The Diffusion Equation

In the previous section, we showed that the departure processes of a GI/G/1 system under heavy traffic conditions and with two classes of customers are renewal processes. The (mean, variance) of the interdeparture times of class 1 and class 2 are $(\mu_{d_1}^{-1}, \sigma_{d_1}^2)$ and $(\mu_{d_2}^{-1}, \sigma_{d_2}^2)$, respectively.

Now let us proceed to derive the diffusion equation for this system.

We first introduce the following notations:

$A_1(t)$ represents the cumulative number of arrivals of class 1 customers up to time t ,

$D_1(t)$ represents the cumulative number of departures of class 1 customers up to time t ,

and $Q_1(t)$ represents the number of customers in the system of class 1 at time t (including the one in service, if any).

Then, $Q_1(t) \stackrel{\Delta}{=} A_1(t) - D_1(t)$.

The change in the number of customers of class 1 in the system between time t and $t + dt$ is

$$\begin{aligned}\Delta Q_1(t) &\stackrel{\Delta}{=} Q_1(t + dt) - Q_1(t) \\ &= \Delta A_1(t) - \Delta D_1(t)\end{aligned}\tag{4.15}$$

where $\Delta A_1(t) \stackrel{\Delta}{=} A_1(t + dt) - A_1(t)$

and $\Delta D_1(t) \stackrel{\Delta}{=} D_1(t + dt) - D_1(t)$

The interarrival times of class 1 customers are i.i.d. with the (mean, variance) given by $(\lambda_1^{-1}, \sigma_{a_1}^2)$. If we assume dt is sufficiently large for many arrivals and departures to take place, and if $Q_1(t)$ does not go to zero in this interval, then from the central-limit theorem $\Delta Q_1(t)$ is approximately normally distributed with the following mean and variance:

$$E[\Delta Q_1(t)] = (\lambda_1 - \mu_{d_1}) dt \tag{4.16}$$

$$\text{var}[\Delta Q_1(t)] = (c_{a_1} \lambda_1^2 + c_{d_1} \mu_{d_1}^2) dt \tag{4.17}$$

where $c_{a_1} = \sigma_{a_1}^2 \lambda_1^2$

and $c_{d_1} = \sigma_{d_1}^2 \mu_{d_1}^2$

Similarly for class 2 customers, $\Delta Q_2(t)$ is approximately normally distributed with the following mean and variance:

$$E[\Delta Q_2(t)] = (\lambda_2 - \mu_{d_2})dt \quad (4.18)$$

$$\text{var}[\Delta Q_2(t)] = (c_{a_2}\lambda_2 + c_{d_2}\mu_{d_2})dt \quad (4.19)$$

Since the arrivals of the two classes are assumed independent of each other, and they are also independent of departures because of the heavy traffic conditions, then

$$\begin{aligned} \text{cov}[\Delta Q_1(t), \Delta Q_2(t)] &= \text{cov}[\Delta D_1(t), \Delta D_2(t)] \\ &= r \sqrt{\text{var}(\Delta D_1(t)) \times \text{var}(\Delta D_2(t))} \\ &= r \sqrt{c_{d_1}\mu_{d_1} c_{d_2}\mu_{d_2}} dt \end{aligned} \quad (4.20)$$

where r is the "infinitesimal" correlation coefficient of the normally distributed variables $\Delta Q_1(t)$ and $\Delta Q_2(t)$, and $|r| \leq 1$

The value of r will be discussed later in this chapter.

Let $\Delta Q(t)$ be a 2-dimensional process

$$\Delta Q(t) \triangleq \begin{bmatrix} \Delta Q_1(t) \\ \Delta Q_2(t) \end{bmatrix} \quad (4.21)$$

If we approximate the 2-dimensional discrete-state process $Q(t)$ by a 2-dimensional continuous process $X(t)$, then $X(t)$ is defined by the stochastic differential equation

$$d\bar{X}(t) = b dt + a^{1/2} dW(t) \quad (4.22)$$

where $W(t)$ is a 2-dimensional Wiener Process with zero mean and unit covariance matrix

b is the 2-dimensional vector of means

$$b = \begin{bmatrix} \lambda_1 - \mu_{d_1} \\ \lambda_2 - \mu_{d_2} \end{bmatrix} \quad (4.23)$$

a is the positive definite covariance matrix

$$a = \begin{bmatrix} \lambda_1 c_{a_1} + c_{d_1} \mu_{d_1} & \sqrt{c_{d_1} \mu_{d_1} c_{d_2} \mu_{d_2}} \\ \sqrt{c_{d_1} \mu_{d_1} c_{d_2} \mu_{d_2}} & \lambda_2 c_{a_2} + c_{d_2} \mu_{d_2} \end{bmatrix} \quad (4.24)$$

Given the initial state of x_0 at $t = 0$, then the conditional probability density at time t ($p(\bar{x}, t)$) satisfies the 2-dimensional diffusion equation with constant coefficients given as

$$\begin{aligned} \frac{\partial}{\partial t} p(\bar{x}, t) = & \frac{1}{2} \sum_{j=1}^2 \sum_{i=1}^2 a_{ij} \frac{\partial^2}{\partial x_j \partial x_i} p(\bar{x}, t) \\ & - \sum_{i=1}^2 b_i \frac{\partial}{\partial x_i} p(\bar{x}, t) \end{aligned} \quad (4.25)$$

4.4 Boundary Conditions and Steady State Solutions

Since $X_1(t)$ represents the number of customers of class 1 and $X_2(t)$ represents the number of customers of class 2 in the system at time t , then both variables will have non negative values. We can ensure this by having instantaneous reflecting boundaries at $x_1 = 0$ and $x_2 = 0$.

As we have shown in Chapter 2 this can be done by letting the directional fluxes ϕ_1 and ϕ_2 equal to zero at $x_1 = 0$ and $x_2 = 0$, respectively.

Thus

$$-\frac{1}{2} a_{11} \frac{\partial p(\bar{x}, t)}{\partial x_1} - \frac{1}{2} a_{12} \frac{\partial p(\bar{x}, t)}{\partial x_2} + b_1 p(\bar{x}, t) = 0 \bigg|_{x_1=0} \quad (4.26)$$

$$\text{and } -\frac{1}{2} a_{12} \frac{\partial p(\bar{x}, t)}{\partial x_1} - \frac{1}{2} a_{22} \frac{\partial p(\bar{x}, t)}{\partial x_2} + b_2 p(\bar{x}, t) = 0 \bigg|_{x_2=0} \quad (4.27)$$

A sufficient condition to satisfy these boundary conditions at $x_1 = 0$ and $x_2 = 0$ is to satisfy both of them everywhere in the (x_1, x_2) space. In fact, we will show later in this section that satisfying both Eqs. 4.26 and 4.27 everywhere in the (x_1, x_2) space will lead to a product form solution for the equilibrium joint density function.

Let us first give an explanation of Eqs. 4.26 and 4.27. If we multiply Eq. 4.26 by b_2 and Eq. 4.27 by b_1 and subtract the first equation from the second, we find that

$$(a_{11} b_2 - a_{12} b_1) \frac{\partial p(\bar{x}, t)}{\partial x_1} - (-a_{12} b_2 + a_{22} b_1) \frac{\partial p(\bar{x}, t)}{\partial x_2} = 0 \quad (4.28)$$

This equation shows that the derivative of $p(\bar{x}, t)$ in the direction of the vector $((a_{11}b_2 - a_{12}b_1), -(-a_{12}b_2 + a_{22}b_1))$ in the (x_1, x_2) space vanishes. In other words, this is the direction of the contours corresponding to constant values of $p(\bar{x}, t)$ in the (x_1, x_2) space.

It will be shown later that the steady state solution of $p(\bar{x}, t)$ exists if the values of $(a_{22}b_1 - a_{12}b_2)$ and $(a_{11}b_2 - a_{12}b_1)$ are negative. Therefore,

$$\tan \eta = - \frac{a_{22}b_1 - a_{12}b_2}{a_{11}b_2 - a_{12}b_1} \quad (4.29)$$

is also negative, and since this slope is constant in the (x_1, x_2) space, the contours are parallel lines with this negative slope as shown in fig. 4.1.

Now we proceed to solve the diffusion equation. Its steady state solution can be obtained by solving the equation

$$\frac{\partial \phi_1\{x_1|x_2\}}{\partial x_1} = - \frac{\partial \phi_2\{x_2|x_1\}}{\partial x_2} \quad (4.30)$$

$$\text{subject to } \phi_1\{x_1|x_2\} \Big|_{x_1=0} = \phi_2\{x_2|x_1\} \Big|_{x_2=0} = 0 \quad (4.31)$$

where

$$\phi_1\{x_1|x_2\} = \phi_1\{x_1|x_2; \infty\}$$

and

$$\phi_2\{x_2|x_1\} = \phi_2\{x_2|x_1; \infty\}$$

As we have shown in Chapter 3, the solution of Eq. 4.30 subject to Eq. 4.31 is equivalent to solving the following simultaneous equations

Fig. 4.1 Contour Lines

$$\phi_1\{x_1|x_2\} = \phi_2\{x_2|x_1\} = 0 \quad (4.32)$$

Therefore, the steady state solution can be obtained by solving the following two simultaneous partial differential equations of first order:

$$\frac{a_{11}}{2} \frac{\partial p}{\partial x_1} + \frac{a_{12}}{2} \frac{\partial p}{\partial x_2} - b_1 p = 0 \quad (4.33)$$

$$\frac{a_{12}}{2} \frac{\partial p}{\partial x_1} + \frac{a_{22}}{2} \frac{\partial p}{\partial x_2} - b_2 p = 0 \quad (4.34)$$

where $p = p(\bar{x}, \infty)$

$$\text{Let } \psi = \log p, \text{ i.e. } p = e^\psi \quad (4.35)$$

$$\text{Then,} \quad a \frac{\partial \psi}{\partial \bar{x}} = 2b \quad (4.36)$$

where a is a positive definite matrix,

b is a vector,

and $\frac{\partial \psi}{\partial \bar{x}}$ is the vector of derivatives $\left\{ \frac{\partial \psi}{\partial x_1}, \frac{\partial \psi}{\partial x_2} \right\}$.

The solution of Eq. 4.36 is

$$\psi = 2(a^{-1} b)^T \bar{x} = \gamma \bar{x} \quad (4.37)$$

where γ is a row vector.

If all the components of the vector γ are negative, the solution

$$p = C \exp(\gamma \bar{x}) \quad (4.38)$$

will satisfy the equation

$$\int_0^{\infty} \int_0^{\infty} p \, dx_1 \, dx_2 = 1$$

which leads to a value of the constant C given as

$$C = \gamma_1 \gamma_2$$

Therefore, the joint density function exists and is given by:

$$p(\bar{x}) = \prod_{i=1}^2 |\gamma_i| \exp\{\gamma_i x_i\} \quad (4.39)$$

which is a product of the following two marginal density functions:

$$p(x_1) = |\gamma_1| \exp\{\gamma_1 x_1\} \quad (4.40)$$

and

$$p(x_2) = |\gamma_2| \exp\{\gamma_2 x_2\} \quad (4.41)$$

It is interesting to show that Eq. 4.39 satisfies the contour lines of fig. 3.1. The proof is given in the following:

Since
$$p(\bar{x}) = C \exp\{\gamma_1 x_1 + \gamma_2 x_2\}$$

and the contour lines are obtained by letting

$$p(\bar{x}) = \text{const.}$$

Therefore, the contour lines must satisfy the following equation:

$$\gamma_1 x_1 + \gamma_2 x_2 = \text{const.} \quad (4.42)$$

which is the equation of a set of parallel lines with slope

$$\frac{dx_2}{dx_1} = -\frac{\gamma_1}{\gamma_2} = -\frac{a_{22}b_1 - a_{12}b_2}{a_{11}b_2 - a_{12}b_1} \quad (4.43)$$

This agrees with Eq. 4.29.

By discretizing $p(x_1)$, we obtain the probability distribution of the number of the customers in the system of class i (n_i) as

$$\begin{aligned} \hat{p}(n_i) &= \int_{n_i}^{n_i+1} p(x_i) dx_i \\ &= (1 - \hat{\rho}_i) \hat{\rho}_i^{n_i} \quad i = 1 \text{ or } 2 \end{aligned} \quad (4.44)$$

where $\hat{\rho}_i = \exp \gamma_i$ (4.45)

Now, we calculate γ from the vector $b = \{b_i\}$ and the matrix $a = \{a_{ij}\}$ as

$$\gamma^T = 2a^{-1} b = \frac{2}{a_{11}a_{22} - a_{12}^2} \begin{bmatrix} (\lambda_1 - \mu_{d1})a_{22} - (\lambda_2 - \mu_{d2})a_{12} \\ -(\lambda_1 - \mu_{d1})a_{12} + (\lambda_2 - \mu_{d2})a_{11} \end{bmatrix} \quad (4.46)$$

We notice that, $\frac{\lambda_1}{\mu_{d1}} = \alpha \lambda \left(\frac{\alpha \mu_1^{-1} + (1 - \alpha) \mu_2^{-1}}{\alpha} \right)$

$$= \rho \left(\alpha + \frac{(1 - \alpha)}{\beta} \right) = \rho_{\text{eff}} \quad (4.47)$$

where $\rho = \lambda / \mu_1$

Similarly, $\frac{\lambda_2}{\mu_{d2}} = \rho_{\text{eff}} \quad (4.48)$

In fact ρ_{eff} is the total utilization of the server by the two classes and is equal to

$$\rho_{\text{eff}} = \rho_1 + \rho_2 = \frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2} \quad (4.49)$$

After some algebraic manipulation, we find

$$\gamma_1 = \frac{-2(1 - \rho_{\text{eff}})}{\rho_{\text{eff}} c_{a_1} + c_{d_2}} \times D_1 \quad (4.50)$$

$$\gamma_2 = \frac{-2(1 - \rho_{\text{eff}})}{\rho_{\text{eff}} c_{a_2} + c_{d_1}} \times D_2 \quad (4.51)$$

where

$$D_1 = \frac{c_{a_2} \rho_{\text{eff}} + c_{d_2} - r \sqrt{c_{d_1} c_{d_2} \left(\frac{1-a}{a} \right)}}{c_{a_2} \rho_{\text{eff}} + c_{d_2} - \frac{r^2 c_{d_1} c_{d_2}}{c_{a_1} \rho_{\text{eff}} + c_{d_1}}} \quad (4.52)$$

and

$$D_2 = \frac{c_{a_1} \rho_{\text{eff}} + c_{d_1} - r \sqrt{c_{d_1} c_{d_2} \left(\frac{a}{1-a} \right)}}{c_{a_1} \rho_{\text{eff}} + c_{d_1} - \frac{r^2 c_{d_1} c_{d_2}}{c_{a_2} \rho_{\text{eff}} + c_{d_2}}} \quad (4.53)$$

The factors D_1 and D_2 represent the effect of the correlation between the two departure processes.

If the output processes are not correlated ($r = 0$), then

$$D_1 = D_2 = 1.$$

* Also note that if $\alpha = 1$, then c_{d_1} is equal to the squared coefficient of variation of service c_1 , and γ_1 will take a form similar to that derived by Kobayashi [Ko74] for the single class queueing system.

From the probability distribution of the number of customers in the system of class i given by Eq. 4.44, we can obtain the mean $E[\hat{n}_i]$ and the variance $\text{var}[\hat{n}_i]$, for $i = 1, 2$, as follows.

$$E[\hat{n}_i] = \hat{\rho}_i / (1 - \hat{\rho}_i) \quad (4.54)$$

$$\text{var}[\hat{n}_i] = \hat{\rho}_i / (1 - \hat{\rho}_i)^2 \quad (4.55)$$

We define the relative error in the mean $E[\hat{n}_i]$ and the variance $\text{var}[\hat{n}_i]$, for $i = 1, 2$, as

$$ER_i = \frac{E[n_i] - E[\hat{n}_i]}{E[n_i]} \times 100 \quad (4.56)$$

$$ERV_i = \frac{\text{var}[n_i] - \text{var}[\hat{n}_i]}{\text{var}[n_i]} \times 100 \quad (4.57)$$

where $E[n_i]$ and $\text{var}[n_i]$ are the values of the mean and the variance, respectively, obtained from analytical expressions or simulation.

Also the relative error in the total number of customers in the system is given as

$$ER = \frac{E[n] - E[\hat{n}]}{E[\hat{n}]} \times 100 \quad (4.58)$$

where

$$E[\hat{n}] = E[\hat{n}_1] + E[\hat{n}_2] \quad (4.59)$$

Correlation Coefficient to Minimize the Relative Error (ER)

So far, an analytical expression for the "infinitesimal" correlation coefficient r (or the corresponding "infinitesimal" covariance a_{12}) has eluded the author. Therefore, in the following we obtain an empirical expression for r . We varied the value of r in the range $[-1, 1]$ and calculated the corresponding diffusion estimates for the mean number in the system using Eqs. 4.59 and 4.54. We obtained values for the relative error (ER), and plotted ER as a function of r . Two examples of these plots are shown in fig. 4.2, which show a linear relation. We found in all the cases studied that the values of r that result in zero relative error (r^*) are negative and in the range -1 to 0 . The value $r = -1$ corresponds to a symmetrical Markov system ($\beta = 1$, $\alpha = .5$, and $c_{a1} = c_{a2} = c_1 = c_2 = 1$) under heavy traffic conditions ($\rho_{eff} = 1$). We conjectured that the value of r , that minimizes the error in estimating the mean total number of customers in the system, is a function of the average squared coefficient of variation for departures ($c_{d_{av}}$) and the average squared coefficient of variation for arrival ($c_{a_{av}}$),

$$\text{where } c_{d_{av}} = \frac{c_{d1} + c_{d2}}{2}, \text{ and } c_{a_{av}} = \alpha c_{a1} + (1 - \alpha) c_{a2}.$$

The Conjecture

The value of r that minimizes the error in estimating the mean total number of customers in the system (r^*) is negative and proportional to

$$c_{a_{av}} / c_{d_{av}},$$

2

Fig. 4.2 Relative Error (ER) as a Function of Correlation Coefficient r for Two Cases

$$\text{i.e., } r^* = -c_{a,av} / c_{d,av} \text{ and } 1 \geq r^* \geq -1 \quad (4.60)$$

To examine this conjecture, we plotted $(c_{a,av} / c_{d,av})$ vs. r^* in fig. 4.3 and found that a line with slope -1 that passes through the point (0,0) can be fitted to the points. This leads us to conclude that the conjecture is acceptable. The results reported later will give further support to the stated conjecture.

The negative correlation coefficient obtained from Eq. 4.60 can be explained by the following heuristic arguments:

- Since there is only one server attending the two classes, we expect that the increase in the number of departures of one class in the interval dt will correspond to a decrease in the number of departures of the other class.

- The instantaneous reflecting boundaries used in the solution of the diffusion equation lead to the following conditions: $Q_1(t) > 0$ and $Q_2(t) > 0$ for all $t > 0$. However, the heavy traffic condition necessitates that $Q_1(t) + Q_2(t) > 0$ for all $t > 0$. Therefore, the instantaneous reflecting boundaries are sufficient but not necessary for heavy traffic, and will lead to overestimation of the mean number in the system for both classes. The negative correlation coefficient tends to compensate and thus correct such overestimation.

The form of Eq. 4.60 will compensate for the positive error in estimating the mean number in the system for individual classes for $c_{a,av} < 1$ and for $c_{s,av} = \alpha c_1 + (1 - \alpha)c_2 > 1$ (ER_1 is plotted as a function of $c_{s,av}$ and $c_{a,av}$ in figs. 3.22 and 3.26 respectively).

Fig. 4.3 Correlation Coefficient that Minimizes the Relative Error ER as a Function of $\frac{c_{a_{av}}}{c_{d_{av}}}$

4.5 Analysis of Results

4.5.1 Introduction

In this section, we check the accuracy of the diffusion approximation. We first compare the mean number of customers in an M/M/1 system with two classes of customers, with known analytical results. Then, we proceed to discuss the errors in the diffusion estimates for GI/G/1 systems by comparing the mean, the variance, and the distribution of the number in the system with simulation results. Two values for the correlation coefficient (r) are used in this section; namely $r = -1$ and a value obtained from the empirical expression given by Eq. 4.60.

4.5.2 Accuracy of the Diffusion Approximation for M/M/1 Systems

The mean number of customers of each class in an M/M/1 queueing system are derived in Morse [Mo65] and in Gross and Harris [GH74] and are given by

$$E[n_1] = \rho_1 + \rho_1 \rho \frac{1 - (1 - \frac{1}{\beta})\rho_2}{1 - \rho_{\text{eff}}} \quad (4.61)$$

$$E[n_2] = \rho_2 + \rho_2 \rho \beta \frac{\frac{1}{\beta^2} + (1 - \frac{1}{\beta})\rho_1/\beta}{1 - \rho_{\text{eff}}} \quad (4.62)$$

$$E[n] = \rho_{\text{eff}} + \frac{\rho\rho_1 + \rho\rho_2/\beta}{1 - \rho_{\text{eff}}} \quad (4.63)$$

where

$$\begin{aligned} \rho_1 &= \lambda_1/\mu_1, \quad \rho_2 = \lambda_2/\mu_2 \\ \rho &= \lambda/\mu_1, \quad \rho_{\text{eff}} = \rho_1 + \rho_2 \\ \alpha &= \lambda_1/\lambda, \quad \beta = \mu_2/\mu_1 \end{aligned}$$

It is worth noting that if $\beta = 1$, then $\rho = \rho_{\text{eff}}$ and $E[n] = \frac{\rho_{\text{eff}}}{1 - \rho_{\text{eff}}}$

which is the standard M/M/1 result. Furthermore, it is easy to prove [Mo65] that for β less than or greater than 1, the mean number of customers in the system is larger for the same effective utilization (ρ_{eff}) (i.e. the minimum queue length corresponds to $\beta = 1$).

The mean number of customers in the system of class i is calculated from the diffusion approximation as

$$E[\hat{n}_i] = \hat{\rho}_i / (1 - \hat{\rho}_i), \quad i = 1, 2$$

where, $\hat{\rho}_i$ is calculated from the expressions derived in the previous section (Eqs. 4.45, 4.50 and 4.51), which we rewrite as

$$\hat{\rho}_i = \exp \left\{ \frac{-2(1 - \rho_{\text{eff}})}{\rho_{\text{eff}} c_{a_i} + c_{d_i}} \times D_i \right\}, \quad i = 1, 2 \quad (4.64)$$

where D_i is a function of the arrival and the departure parameters and of the correlation coefficient between the number of departures of the two classes in time intervals dt as shown in Eqs. 4.52 and 4.53. In these expressions, we assume r takes two values; namely $r = -1$ and r calculated from Eq. 4.60.

We first discuss the case of $r = -1$. Figures 4.4 and 4.5 show the relative error in the mean total number of customers in the system of both classes (ER), defined by Eq. 4.58, as a function of α for different values of ρ_{eff} and β .

Fig. 4.4 Relative Error (ER) as a Function of α for $\beta = 1$ and Different Values of ρ_{eff}

Fig. 4.5 Relative Error (ER) as a Function of α for $\rho_{\text{eff}} = 0.9$ and Different Values of β

We observe the following:

- The relative error is always positive (underestimating the mean).
- The relative error decreases as $\rho \rightarrow 1$, which is expected from the heavy traffic conditions.
- For $\beta = 1$, the minimum error is always at $\alpha = .5$, though the range of changes in error with α is relatively narrow.
- As β deviates from 1, the minimum error occurs at $\alpha < .5$ for $\beta > 1$ and at $\alpha > .5$ for $\beta < 1$.

Figures 4.6 and 4.7 show the relative error in the mean number of customers in the system of class 1 as a function of α , β and ρ_{eff} . The relative error here changes sign. We see that for a symmetrical system ($\alpha = .5$ and $\beta = 1$) the error is positive and approaches zero as $\rho_{\text{eff}} \rightarrow 1$ (dotted curves). Also, we note that changing α has a similar effect on the relative error as changing β . The error is positive and increases as we increase α from .5 or β from 1. For these values of α and β , the utilization of the server by class 1 customers (ρ_1) is higher than that by class 2 customers (ρ_2), and the assumption that $r = -1$ tends to underestimate the class with high utilization and overestimate that with low utilization (see the dotted curves in figs. 4.6 and 4.7).

We expect that if we plot ER_1 vs. α for different values of β , the crossing point will be at $\alpha < .5$ for $\beta > 1$ and at $\alpha > .5$ for $\beta < 1$. In fact the crossing point will be around a point that makes ρ_1 equal to ρ_2 ,

$$\text{i.e., } \frac{\lambda_1}{\mu_1} = \frac{\lambda_2}{\mu_2} \quad \text{or} \quad \frac{\alpha}{1-\alpha} \beta = 1 \quad (4.65)$$

This is true for a symmetrical system, and figs. 4.8 and 4.9 show that it is also valid for nonsymmetrical systems.

Fig. 4.6 Relative Error in the Mean Number in the System of Class 1 as a Function of α for $\beta = 1$ and Different Values of ρ_{eff}

Fig. 4.7 Relative Error in the Mean Number in the System of Class 1 as a Function of β for $\alpha = 0.5$ and different Values of ρ_{eff}

Fig. 4.8 Relative Error in the Mean Number in the System of Class 1 as a Function of β for $\rho_{eff} = 0.95$ and Different Values of α

Fig. 4.9 Relative Error in the Mean Number in the System of Class 1 as a Function of α for $\rho_{eff} = 0.95$ and Different Values of β

Finally, we note that the relative errors in the diffusion approximation for the mean number of customers of one class saturates as this class becomes dominant (i.e. $\alpha \rightarrow 1$ or $\rho \gg 1$).

We now study the effect of using r^* obtained from Eq. 4.60. Figures 4.10 and 4.11 show comparison between the results for $r = -1$ and r^* .

From these figures, we have the following observations:

- If $\beta = 1$ then $r = r^* = -1$ for all values of α . Therefore, there is no change in the accuracy.
- For other values of β , there is improvement in the accuracy of the estimate of the mean total number in the system for all values of α as shown in fig. 4.10. However, for the individual classes, the improvement is only for the positive error (underestimation) as shown in fig. 4.11. Therefore, we conclude that the estimates with r^* are better because they are 'pessimistic' compared to those with $r = -1$. With r^* , the underestimation of the queue length is decreased for some values of α and β at the expense of an increase in the overestimation of the queue length for other values of α and β .

4.5.3 Accuracy of the Diffusion Approximation for GI/G/1 Systems

In this section, we compare the diffusion approximations of the mean, the variance, and the distribution of the number in the system of class i with the corresponding values obtained from GPSS simulation. Results are presented with a discussion of the effect of changing the parameters

Fig. 4.10 Comparison Between the Relative Error ER
for $r = -1$ and r^* ($\rho_{eff} = 0.9$)

Fig. 4.11 Comparison Between the Relative Error in Estimates of Class 1 for $r = -1$ and r^* ($\rho_{\text{eff}} = 0.95$)

α , β , ρ_{eff} , the squared coefficient of variation of the service time (c_s), and the squared coefficient of variation of the interarrival time (c_a) on the accuracy of the approximations. Two values of r ($r = 1$ and r^* obtained from Eq. 4.60) are used in this section.

Simulation

Simulation of the queueing system is performed using the GPSS language, since it is very appropriate for simulating queueing systems. The program consists of about 50 blocks and was run for 10,000 or 20,000 transactions according to the utilization of the server. (A GPSS transaction represents a customer in the queueing system).

The runs were performed on the Carleton University XDS SIGMA-9 computer and the CPU time was between 2 to 4 minutes, where the smaller value corresponds to the smaller number of transactions. The flowchart of the simulation program is presented in Appendix B.

Numerical Results

In this section, we compare the results obtained from the diffusion approximation to those obtained from the simulation. The comparison is done for the distribution of the number of customers in the system of each class and the corresponding mean and variance. We study the accuracy of the approximation as a function of the following parameters:

- the arrival ratio ($\alpha = \frac{\lambda_1}{\lambda}$)
- the ratio of service rates ($\beta = \frac{\mu_2}{\mu_1}$)
- the effective utilization ($\rho_{\text{eff}} = \rho_1 + \rho_2$)
- the service squared coefficients of variation (c_1 and c_2)
- the arrival squared coefficients of variation (c_{a_1} and c_{a_2})

For changing c_1 and c_2 , we chose the following:

- Two stages with $\pi_1 = \pi_2 = 1$ (i.e. Erlang 2 or E_2).

Hence, $c_1 = c_2 = \frac{1}{2}$ if both stages have the same rate of service.

- One stage with $\pi_1 = 1$ (simple exponential E_1).

Hence, $c_1 = c_2 = 1$.

- Two stages with general π_1 and π_2 (H_2). The values of c_1 and c_2 are functions of the probabilities π_1 and π_2 , and the service rates μ_1 and μ_2 . This service facility is a hyperexponential one.

Similarly, for c_a , we chose interarrival times with uniform distribution ($c_a = .33$), Erlang-2 distribution ($c_a = .5$), exponential distribution ($c_a = 1$), and hyperexponential distribution ($c_a = 2$) where $c_{a1} = c_{a2} = c_a$.

As we mentioned before the mean number of customers in the system is calculated from the diffusion approximation as

$$E[\hat{n}_i] = \hat{\rho}_i / (1 - \hat{\rho}_i) \quad , \quad i = 1, 2$$

The variance is calculated as

$$\text{var}[\hat{n}_i] = \hat{\rho}_i / (1 - \hat{\rho}_i)^2 \quad , \quad i = 1, 2$$

and the distribution as

$$p(\hat{n}_i) = (1 - \hat{\rho}_i) \hat{\rho}_i^{n_i} \quad , \quad i = 1, 2$$

where $\hat{\rho}_i$ is calculated from Eq. 4.64 with r given two values (i.e. $r = -1$ and r^* obtained from Eq. 4.60).

Now, we proceed with some selected cases. Other unreported cases have shown results consistent with the observations we make in the remainder of this chapter.

2-Stage Service Facility (H_2)

A number of simulation runs were performed to test the accuracy of the diffusion approximation for a queueing system with squared coefficient of variation for service time greater than 1. The other parameters of the queueing system (α , β and ρ_{eff}) were changed to study the sensitivity of the accuracy to these parameters. The number of cases studied were limited due to the high cost of simulation runs.

Tables 4.1 and 4.2, and figs. 4.12 and 4.13 show the effect of changing β , for $\alpha = .5$ and $\rho_{\text{eff}} = .91$. This effect is also shown in fig. 4.17. Tables 4.3 and 4.4, and figs. 4.14 and 4.15 are for a queueing system with $\alpha = .5$ and $\rho_{\text{eff}} = .82$ and different values of β . This is also plotted in fig. 4.17. From fig. 4.17, we see that the error in estimating the mean number in the system for $\alpha = .5$ is minimum if $\beta = 1$, and increases as we deviate from this value. This is again due to the departure from the values of α and β of the "symmetrical system", i.e. $\alpha = .5$ and $\beta = 1$, which we discussed in the M/M/1 cases. We notice also that the error is less for the systems with high utilization, which is expected due to the nature of the diffusion approximation (good approximation for heavy traffic). In general, we find the errors in estimating the variances are higher than the corresponding errors in estimating the means. Table 4.5 and fig. 4.16 give the results for $\alpha = .74$, and fig. 4.18 summarizes the effect of changing α on the accuracy of the diffusion estimates. We notice the same trends that appeared in the M/M/1 cases which are: (1) underestimating queue

Poisson arrival

 $\alpha = .5$

2-stage service facility ($\mu_1^{-1} = \mu_2^{-1} = 24$)
 ($c_1 = c_2 = 1.583$)

 $\beta = 1$ $\rho_{\text{eff}} = .91$ $c_{d1} = c_{d2} = 1.29$

		Queue 1	Queue 2	Total
<u>Sim</u>	mean	5.69	5.73	11.42
	90% Conf. Int. for mean	[4.69-6.47]	[4.74-6.51]	
	var	38.1	37.7	
<u>Diff</u>	$\hat{\rho}_1$	.828	.828	
$r = -1$	mean	4.83	4.83	9.66
	ER_1	15.1	15.7	15.4
	var	28.1	28.1	
	ERV_1	26.2	25.5	
<u>Diff</u>	$\hat{\rho}_1$	.864	.864	
$r = -.8$	mean	6.33	6.33	12.66
	ER_1	-11.2	-10.5	-10.8
	var	46.5	46.5	
	ERV_1	-22	-23.3	

Table 4.1 Accuracy of the Diffusion Approximation for Case 1

Fig. 4.12 Marginal Queue Length Distributions for Case 1

Poisson arrival

 $\alpha = .5$ 2-stage service facility ($\mu_1^{-1} = 12, \mu_2^{-1} = 24$) $\beta = .5$ $(c_1 = c_2 = 1.583)$ $\rho_{\text{eff}} = .91$ $c_{d_1} = 1.77, c_{d_2} = 1.1$

		Queue 1	Queue 2	Total
<u>Sim</u>	mean	6.62	7.05	13.67
	90% Conf. Int.	[4.86-7.91]	[5.15-8.42]	
	for mean			
	var	47.16	55.22	
<u>Diff</u>	$\hat{\rho}_1$	.833	.804	
$r = -1$	mean	4.99	4.1	9.09
	ER_1	24.7	41.8	33.7
	var	29.88	20.92	
	ERV_1	36.6	62	
<u>Diff</u>	$\hat{\rho}_1$	.88	.86	
$r = -.7$	mean	7.56	6.09	13.65
	ER_1	-14	13.6	0.1
	var	63	43.5	
	ERV_1	-33.5	21.2	

Table 4.2 Accuracy of the Diffusion Approximation for Case 2

Fig. 4.13 Marginal Queue Length Distributions for Case 2

Poisson arrival

 $\alpha = .5$

2-stage service facility ($\mu_1^{-1} = \mu_2^{-1} = 24$)
 ($c_1 = c_2 = 1.583$)

 $\beta = 1$ $\rho_{\text{eff}} = .82$ $c_{d1} = c_{d2} = 1.29$

		Queue 1	Queue 2	Total
<u>Sim</u>	mean	2.55	2.66	5.21
	90% Conf. Int.	[2.26-2.79]	[2.33-2.93]	
	for mean			
	var	7.35	9.53	
<u>Diff</u> $r = -1$	$\hat{\rho}_1$	.652	.652	
	mean	1.88	1.88	3.76
	ER_1	26.3	29.3	27.8
	var	5.4	5.4	
	ERV_1	26.5	43.3	
<u>Diff</u> $r = -.8$	$\hat{\rho}_1$	.722	.722	
	mean	2.6	2.6	5.2
	ER_1	-2.0	2.3	0.2
	var	9.35	9.35	
	ERV_1	-27.2	1.9	

Table 4.3 Accuracy of the Diffusion Approximation for Case 3

Fig. 4.14 Marginal Queue Length Distributions for Case 3

Poisson arrival

 $\alpha = .5$ 2-stage service facility ($\mu_1^{-1} = 12, \mu_2^{-1} = 24$) $\beta = .5$ $(c_1 = c_2 = 1.583)$ $\rho_{\text{eff}} = .82, \quad c_{d_1} = 1.77, \quad c_{d_2} = 1.10$

		Queue 1	Queue 2	Total
<u>Sim</u>	mean	2.84	2.99	5.83
	90% Conf. Int.	[2.32-3.29]	[2.54-3.37]	
	for mean			
	var	15.74	13.26	
<u>Diff</u>	$\hat{\rho}_i$	.674	.622	
$r = -1$	mean	2.07	1.65	3.72
	ER_i	27.1	44.8	36.2
	var	6.35	4.37	
	ERV_i	59.6	67	
<u>Diff</u>	$\hat{\rho}_i$	.746	.699	
$r = -.8$	mean	2.93	2.32	5.26
	ER_i	-3.2	22.4	9.8
	var	11.54	7.71	
	ERV_i	26.7	41.8	

Table 4.4 Accuracy of the Diffusion Approximation for Case 4

Fig. 4.15 Marginal Queue Length Distributions for Case 4

Poisson arrival

 $\alpha = .74$

2-stage service facility ($\mu_1^{-1} = \mu_2^{-1} = 24$)
 ($c_1 = c_2 = 1.583$)

 $\beta = 1$

$\rho_{\text{eff}} = .91$ $c_{d1} = 1.43$, $c_{d2} = 1.15$

		Queue 1	Queue 2	Total
<u>Sim</u>	mean	9.63	3.38	13.01
	90% Conf. Int.	[7.2-11.47]	[2.5-4.06]	
	for mean			
	var	111.3	16.7	
<u>Diff</u>	$\hat{\rho}_i$	.842	.76	
$r = -1$	mean	5.34	3.17	8.51
	ER_i	44.5	6.2	34.6
	var	33.8	13.21	
	ERV_i	69.6	20.9	
<u>Diff</u>	$\hat{\rho}_i$	.872	.812	
$r = -.8$	mean	6.83	4.31	11.1
	ER_i	29.1	-27.5	15.4
	var	53.4	22.9	
	ERV_i	52	-37	

Table 4.5 Accuracy of the Diffusion Approximation for Case 5

Fig. 4.16 Marginal Queue Length Distributions for Case 5

Fig. 4.17 Accuracy of Diffusion Approximation for the Mean Number in the System of Class 1 as a Function of β for Different Values of ρ_{eff} , with $\alpha = 0.5$, $c_1 = c_2 = 1.583$, and $r = -1$

Fig. 4.18 Effect of α on the Accuracy of the Diffusion Approximation of the Mean Number in the System of Class 1 Customers for $\beta = 1$, $\rho_{eff} = 0.91$, and $r = -1$

lengths for $\alpha > .5$, (2) the percentage of error increases with α , and (3) as α decreases, we expect a change of the sign of the error, i.e. there will be an overestimation for the class with the small percentage in the system as has been shown in fig. 4.6.

Accuracy vs. $c_{s_{av}}$

In order to study the effect of changing the squared coefficients of variation of service times (c_1 and c_2) on the accuracy of estimating the mean number of customers in the queueing system, simulation with 2-stage Erlang (E_2), exponential and 2-stage hyperexponential (H_2) were performed. The rest of the parameters of the queueing system were fixed at $\rho_{eff} = .91$, $\alpha = .5$ and $\beta = 1$. The results are presented in table 4.7, table 4.6 and table 4.1, and figs. 4.20, 4.19 and 4.12, respectively. From these tables we plot fig. 4.21 which shows that we tend to underestimate the number in the system if $c_1 = c_2 > 1$ and overestimate if $c_1 = c_2 < 1$.

Accuracy vs. $c_{a_{av}}$

The arrival process was changed from Poisson to renewal with uniform, Erlang-2 or hyperexponential interarrival times. The results are reported in tables 4.8 to 4.10 and figs. 4.22 to 4.24, in addition to table 4.6 and fig. 4.19. From these tables, fig. 4.25 is plotted which shows that the accuracy is satisfactory for $c_{a_1} = c_{a_2} \geq 1$. As for $c_{a_1} = c_{a_2} < 1$, the error is positive (underestimation), and higher than cases for $c_{a_1} = c_{a_2} \geq 1$.

Poisson arrival

 $\alpha = .5$ Exponential service
time distribution $(\mu_1^{-1} = \mu_2^{-1} = 24)$
 $(c_1 = c_2 = 1)$ $\beta = 1$ $\rho_{\text{eff}} = .91$ $c_{d1} = c_{d2} = 1$

		Queue 1	Queue 2	Total
<u>Sim</u>	mean	5.27	5.27	10.54
	90% Conf. Int.	[3.83-6.34]	[3.82-6.25]	
	for mean			
	var	31.17	30.37	
<u>Diff</u>	$\hat{\rho}_i$	.834	.834	
$r = -1$	mean	5.04	5.04	10.08
	ER_i	4.36	3.45	4.36
	var	30.36	30.36	
	ERV_i	2.6	.03	

Table 4.6 Accuracy of the Diffusion Approximation for Case 6

Fig. 4.19 Marginal Queue Length Distributions for Case 6

Poisson arrival

 $\alpha = .5$ 2-stage Erlang (E_2)
service time
distribution $(\mu_1^{-1} = \mu_2^{-1} = 24)$ $\beta = 1$ $(c_1 = c_2 = .5)$ $\rho_{\text{eff}} = .91$ $c_{d1} = c_{d2} = .75$

		Queue 1	Queue 2	Total
<u>Sim</u>	mean	4.07	4.07	8.14
	90% Conf. Int.	[2.94-4.96]	[2.89-5.01]	
	for mean			
	var	21.59	24.11	
<u>Diff</u>	$\hat{\rho}_i$	.807	.807	
$r = -1$	mean	4.17	4.17	8.34
	ER_i	-2.46	-2.46	-2.46
	var	21.6	21.6	
	ERV_i	-.05	10.4	

Table 4.7 Accuracy of the Diffusion Approximation for Case 7

Fig. 4.20 Marginal Queue Length Distributions for Case 7

Fig. 4.21 Effect of Changing the Squared Coefficient of Variation of Service Time on the Accuracy of Estimating the Mean Number in the System of Class 1 for $\rho_{eff} = 0.91$, $\alpha = 0.5$, and $\beta = 1$

Hyperexponential
interarrival time distribution $(c_{a_1} = c_{a_2} = 2)$ $\alpha = .5$

Exponential service time
distribution $(\mu_1^{-1} = \mu_2^{-1} = 24)$ $\beta = 1$
 $(c_1 = c_2 = 1)$ §

$\rho_{\text{eff}} = .91$ $c_{d_1} = c_{d_2} = 1$

		Queue 1	Queue 2	Total
<u>Sim</u>	mean	8.53	8.24	16.77
	90% Conf. Int.	[5.25-10.68]	[5.06-10.36]	
	for mean			
	var	98.45	80.87	
<u>Diff</u>	$\hat{\rho}_i$	.9	.9	
$r = -1$	mean	9.04	9.04	18.08
	ER_i	-6	-9.7	-7.8
	var	90.4	90.4	
	ERV_i	8.2	-11.8	

Table 4.8 Accuracy of the Diffusion Approximation for Case 8

Fig. 4.22 Marginal Queue Length Distributions for Case 8

Fig. 4.23 Marginal Queue Length Distributions for Case 9

Fig. 4.24 Marginal Queue Length Distributions for Case 10

Fig. 4.25 Effect of Changing the Squared Coefficient of Variation of Arrival Times on the Accuracy of Estimating the Mean Number in the System of Class 1 for $\rho_{eff} = 0.91$, $\alpha = 0.5$, and $\beta = 1$

4.6 Concluding Remarks

In this chapter, we derived expressions for the diffusion approximation of the distribution of the number of customers in a 2-class GI/G/1 queueing system with FCFS service discipline. Those expressions were tested against analytical formulas and simulation runs. The sensitivity of the accuracy of the estimates to variations in the queueing system parameters (α , β , $c_{s_{av}}$, $c_{a_{av}}$, and ρ_{eff}) were estimated. We found the following observations with the value of $r = -1$:

- Error is the lowest for the symmetrical Markov system ($\alpha = .5$, $\beta = 1$ and $c_a = c_s = 1$) under heavy traffic.
- Error in estimating the mean number of customers in the system of class Y is positive (underestimate) for $c_1 = c_2 \geq 1$, or $c_{a_1} = c_{a_2} \leq 1$, and negative (overestimate) for $c_1 = c_2 < 1$ or $c_{a_1} = c_{a_2} > 1$.
- The relative error ER tends to zero as the total utilization approaches one.
- In estimating the individual queues, we underestimated the class with the higher utilization of the server and overestimated the class with the lower utilization of the server.

We noticed improvement in the accuracy of estimating the total number in the system if we use the relation $r^* = \frac{-c_{a_{av}}}{c_{d_{av}}}$ in Eqs. 4.52

and 4.53. As for estimating the individual queues with r^* , we noticed improvement in most of the cases studied. Finally, the method derived here is not restricted to two-class systems, but can be easily extended to the general multiclass (more than 2) GI/G/1 queueing systems.

5. DIFFUSION APPROXIMATION FOR ALTERNATING PRIORITY QUEUEING SYSTEMS

5.1 Introduction

So far, we discussed two types of service discipline for queues with multiclass, namely FCFS and Preemptive-Resume. In the first discipline there is no priority attached to a particular class of customers, and in the second discipline a priority is assigned to customers by an external mechanism and is independent of the server status. In this chapter, we study a different type of queueing systems in which the priority of a customer is assigned according to the status of the queueing system and the customer's class.

Consider a queueing system with two classes of customers and a single server. Customers of class 1 have priority over customers of class 2 if a class 1 customer is in service, and vice-versa. In other words, if a customer of a certain class arrives while the server is busy serving the other class, then this customer will have to wait until the system is empty of all the customers of the other class and all those of his class who arrived before him. Also, if the server is idle when a customer arrives, then he will acquire priority for his class. Under this discipline, the server is switched between the two classes of customers; however, this switching will not occur until the system is empty of all the customers of the particular class that is receiving service at any time. Because of the switching of priorities, this discipline is called alternating priority discipline.

This discipline is useful in practical situations that involve time and cost in switching service from one class of customers to another,

because it minimizes the number of switching times. It is used to model a manufacturing situation in which a machine is used to service more than one type of item, and switching service between types involves setup-time and setup-cost. In some traffic situations, it is more beneficial to allow all the vehicles flowing in one direction to pass first before changing the traffic direction. Similar to this situation is the model of a half duplex channel connecting two computers [Ch74, Ch75].

The alternating priority (or zero-switch rule) discipline was first studied by Avi-Itzhak et al [AM65]. They obtained expressions for the mean queue length and the mean waiting time for the two classes. Later, Takacs [Ta68] presented a more formal derivation for Avi-Itzhak's results. Recently, Taube-Netto [TN77] extended Takacs' work to the case where customers arrive at two-stage queueing system with one server who switches instantaneously from one stage to the other whenever the number of customers in the present stage is zero.

A related work is the diffusion approximation analysis of a loop service by Fischer [Fi77]. In this discipline the server attends a fixed number of stations in a cyclic pattern and allocates a fixed amount of time for each station during a given cycle. Fischer obtains expressions for the expected waiting time for each station.

In this chapter, we study the instantaneous alternating priority discipline using the diffusion approximations. We start by analysing the departure processes of the queueing system in sec. 5.2. Then, we derive a diffusion equation and obtain its steady state solutions in sec. 5.3. In sec. 5.4, we compare the diffusion results with results obtained from analytical formulas and simulation runs.

5.2 The Departure Process

Assume two classes of customers arrive independently at a single server. Their interarrival times have exponential distributions with means λ_1^{-1} and λ_2^{-1} , respectively. The service time of class i ($i = 1, 2$) has a general distribution function $F_i(x_i)$ with a mean service time μ_i^{-1} , a variance σ_i^2 and a squared coefficient of variation c_i . Service times of class 1 are independent from those of class 2. As in the previous chapters, β is the ratio of service rates (i.e. $\beta = \mu_2/\mu_1$), and α and $1 - \alpha$ are the probabilities of next arrival being of the first class or the second class, respectively, where $\alpha = \lambda_1/\lambda$. The system is under heavy traffic if

$$\lambda_1 \mu_1^{-1} + \lambda_2 \mu_2^{-1} = \rho_1 + \rho_2 = \rho_{\text{eff}} \approx 1$$

For a system with $\rho_{\text{eff}} < 1$, the server undergoes a cycle consisting of two periods, namely the busy period and the idle period. As ρ_{eff} approaches 1, the idle period approaches zero. Hence, for heavy traffic we assume the queue is never empty and the server is always busy with class 1 or class 2 customers.

The time elapsing from the start of service of class i customers who arrived during service of the other class j ($j \neq i$) until the system is empty of class i customers is a r.v. Y_i with mean $\bar{Y}_i$. The number of customers departing in a subperiod Y_i is also a r.v. N_i with mean $\bar{N}_i$.

Figure 5.1 shows a time-diagram of the server with marks indicating the departures of both classes.

Now, we will show that

$$\frac{\bar{N}_1}{\bar{N}_2} = \frac{\lambda_1}{\lambda_2} \quad (5.1)$$

The subperiod Y_1 is the sum of an infinite series of the times T_{li} , ($1 \leq l \leq \infty$), where T_{1i} is the time taken to service customers of class i who arrived during the previous subperiod Y_j ($j \neq 1$), and T_{li} is the time taken to service customers of class i who arrived during the time $T_{l-1,i}$. We now show that the series T_{li} ($1 \leq l \leq \infty$) converges if we consider the mean values of T_{li} . Since the mean of the r.v. Y_1 is given by

$$\bar{Y}_1 = \bar{T}_{11} + \bar{T}_{21} + \bar{T}_{31} + \dots \quad (5.2)$$

where $\bar{T}_{11}$ = the number of customers who arrived during $\bar{Y}_2$

times the mean time of service

$$= \bar{Y}_2 \lambda_1 / \mu_1$$

$$= \bar{Y}_2 \rho_1$$

$\bar{T}_{21}$ = the number of customers who arrived during $\bar{T}_{11}$

times the mean service time

$$= \bar{T}_{11} \lambda_1 / \mu_1 = \bar{Y}_2 \rho_1^2$$

etc.

then,

$$\begin{aligned} \bar{Y}_1 &= \bar{Y}_2 \rho_1 (1 + \rho_1 + \rho_1^2 + \dots) \\ &= \frac{\bar{Y}_2 \rho_1}{1 - \rho_1} \end{aligned} \quad (5.3)$$

But $1 - \rho_1 \approx \rho_2$ because of the heavy traffic assumption,
therefore,

$$\frac{\bar{Y}_1}{\rho_1} = \frac{\bar{Y}_2}{\rho_2} = k \text{ (constant)} \quad (5.4)$$

The average number of departures in $\bar{Y}_1$ is

$$\bar{N}_1 = \bar{Y}_1 \mu_1 = \rho_1 k \mu_1 = \lambda_1 k$$

Similarly $\bar{N}_2 = \lambda_2 k$

Therefore,
$$\frac{\bar{N}_1}{\bar{N}_2} = \frac{\lambda_1}{\lambda_2}$$

It is clear that the distribution of the interdeparture times of class i ($i = 1$ or 2) depend on whether the departure follows a switch of service or not. In fig. 5.1, the departure times of class 1 customers are marked, and we find that the time from I - II, II - III, and IV - V have the distribution of the service time of class 1, while the time III - IV has a different distribution. Therefore, the departure of class 1 customers is not a renewal process. The same argument applies for class 2 customers.

In the following, we present an approximate analysis of the interdeparture times. We replace the departure processes of both classes by "equivalent" renewal processes that have the same rate of departure as the original processes but i.i.d. interdeparture times. This is done as follows: Let Z_i be a r.v. that represents the interdeparture time of the "equivalent" output of class i , and using the average number of departures $\bar{N}_i$ in subperiods Y_i ($i = 1, 2$), then Z_i is given by

x Class 1 departure

• Class 2 departure

Fig. 5.1' Time-Diagram Example for the Alternating Priority Service Discipline

$$\begin{aligned}
 z_i &= \frac{\bar{N}_1 - 1}{\bar{N}_1} x_i + \frac{1}{\bar{N}_1} (\bar{N}_j x_j + x_i) \\
 &= x_i + \frac{\bar{N}_j}{\bar{N}_1} x_j \quad j \neq i \\
 &= x_i + \frac{\lambda_j}{\lambda_i} x_j \quad (5.5)
 \end{aligned}$$

Therefore for class 1, the interdeparture time is given by

$$Z_1 = x_1 + \frac{\lambda_2}{\lambda_1} x_2 \quad (5.6)$$

The mean is given as

$$\begin{aligned}
 \mu_{d_1}^{-1} &= \bar{Z}_1 = E[Z_1] = \mu_1^{-1} + \frac{\lambda_2}{\lambda_1} \mu_2^{-1} \\
 &= \mu_1^{-1} + \frac{1 - \alpha}{\alpha} \mu_2^{-1} \quad (5.7)
 \end{aligned}$$

which is the same mean obtained in the FCFS case.

The variance is given by

$$\sigma_{d_1}^2 = \text{var}[Z_1] = \sigma_1^2 + \left(\frac{1 - \alpha}{\alpha}\right)^2 \sigma_2^2 \quad (5.8)$$

and the squared coefficient of variation is given by

$$\begin{aligned}
 c_{d_1}^2 &= \frac{\sigma_{d_1}^2}{\mu_{d_1}^2} \\
 &= \frac{\alpha^2 c_1 + \left(\frac{1 - \alpha}{\beta}\right)^2 c_2}{\left(\alpha + \frac{1 - \alpha}{\beta}\right)^2} \quad (5.9)
 \end{aligned}$$

Similarly for class 2, we find

$$\mu_{d_2}^{-1} = \bar{z}_2 = \mu_2^{-1} + \frac{\alpha}{1-\alpha} \mu_1^{-1} \quad (5.10)$$

$$\sigma_{d_2}^2 = \sigma_2^2 + \left(\frac{\alpha}{1-\alpha}\right)^2 \sigma_1^2 \quad (5.11)$$

$$c_{d_2} = \frac{\alpha^2 c_1 + \left(\frac{1-\alpha}{\beta}\right)^2 c_2}{\left(\alpha + \frac{1-\alpha}{\beta}\right)^2} = c_{d_1} \quad (5.12)$$

Notice that the average departure rates of the "equivalent" system are equal to those of the original system (also equal to the FCFS rates studied in Chapter 4), while the squared coefficients of variation are approximations due to replacing some r.v.^s by their means in Eq. 5.5.

As discussed before, the results apply to the GI/G/1 system in the limit as $t \rightarrow \infty$ (i.e. in the steady-state).

5.3 The Diffusion Equation and the Steady State Solutions

5.3.1 The Diffusion Equation

As shown in Chapter 4, if the arrival and departure of class i are renewal processes, then the change in the number of customers in small time interval dt ($\Delta Q_i(t)$) is approximately normally distributed with the following mean and variance:

$$E[\Delta Q_i(t)] = (\lambda_i - \mu_{d_i}) dt \quad (5.13)$$

$$\text{and var } [\Delta Q_1(t)] = (c_{a_1} \lambda_1 + c_{d_1} \mu_{d_1}) dt, \quad i = 1, 2 \quad (5.14)$$

The covariance is given by

$$\text{cov } [\Delta Q_1(t), \Delta Q_2(t)] = r \sqrt{c_{d_1} c_{d_2} \mu_{d_1} \mu_{d_2}} \quad (5.15)$$

If $X(t)$ is the 2-dimensional continuous process that approximates the discrete process $Q(t)$, then the conditional probability density at time t given initial conditions $(p(\bar{x}, t))$ satisfies the 2-dimensional diffusion equation

$$\begin{aligned} \frac{\partial}{\partial t} p(\bar{x}, t) = & \frac{1}{2} \sum_{j=1}^2 \sum_{i=1}^2 a_{ij} \frac{\partial^2}{\partial x_j \partial x_i} p(\bar{x}, t) \\ & - \sum_{i=1}^2 b_i \frac{\partial}{\partial x_i} p(\bar{x}, t) \end{aligned} \quad (5.16)$$

5.3.2 Steady State Solutions

Introducing 2 reflecting boundaries at $x_1 = 0$ and $x_2 = 0$, we showed in Chapter 4 that

$$p(\bar{x}) = \prod_{i=1}^2 |\gamma_i| \exp\{\gamma_i x_i\} \quad (5.17)$$

where $\gamma^T = 2a^{-1} b$

and the solution is valid if γ has negative components, i.e. $\rho_{\text{eff}} < 1$.

The discrete marginal distributions are

$$\hat{p}(n_i) = (1 - \hat{\rho}_i) \hat{\rho}_i, \quad i = 1, 2 \quad (5.18)$$

where $\hat{\rho}_i = \exp \gamma_i$ (5.19)

$$\gamma_i = \frac{-2(1 - \rho_{\text{eff}})}{\rho_{\text{eff}} c_{a_i} + c_{d_i}} D_i \quad (5.20)$$

$$\rho_{\text{eff}} = \lambda_1/\mu_1 + \lambda_2/\mu_2 \quad (5.21)$$

$$D_1 = \frac{c_{a_2} \rho_{\text{eff}} + c_{d_2} - r c_{d_1} \sqrt{\frac{1-\alpha}{\alpha}}}{c_{a_2} \rho_{\text{eff}} + c_{d_2} - \frac{r^2 c_{d_1} c_{d_2}}{\rho_{\text{eff}} c_{a_1} + c_{d_1}}} \quad (5.22)$$

$$D_2 = \frac{c_{a_1} \rho_{\text{eff}} + c_{d_1} - r c_{d_2} \sqrt{\frac{\alpha}{1-\alpha}}}{c_{a_1} \rho_{\text{eff}} + c_{d_1} - \frac{r^2 c_{d_1} c_{d_2}}{\rho_{\text{eff}} c_{a_2} + c_{d_2}}} \quad (5.23)$$

$$c_{d_1} = c_{d_2}$$

r is the "infinitesimal" correlation coefficient

$$-1 \leq r \leq 1$$

Correlation Coefficient to Minimize the Relative Error (ER)

For the reasons discussed in Chapter 4, we look for an empirical expression for the correlation coefficient r as follows:

We changed r in the range $[-1, +1]$ and calculated the corresponding diffusion estimates for the mean number in the system (the total of both classes) using Eqs. 5.29 and 5.30. Then, we obtained values for r that make these estimates equal to those calculated from the analytical formulas given by Eqs. 5.27 and 5.28. We found that r^* is always negative and has a value between 0 and -1. The value $r^* = -1$ corresponds to a Symmetrical Markov System ($\beta = 1$, $\alpha = .5$, $c_1 = c_2 = 1$) under heavy traffic conditions ($\rho_{\text{eff}} = 1$). As we varied β , c_1 , c_2 , c_{a1} , and c_{a2} we found that r^* takes different values. Therefore, we conjectured that r^* is a function of β , $c_{s_{\text{av}}} = \alpha c_1 + (1 - \alpha)c_2$, and $c_{a_{\text{av}}} = \alpha c_{a1} + (1 - \alpha)c_{a2}$.

Conjecture

The value r that minimizes the error in estimating the mean total number of customers in the system is negative and is given by

$$r^* = -\min. \left(\beta, \frac{1}{c_{s_{\text{av}}}} \right) \cdot c_{a_{\text{av}}} \quad (5.24)$$

$$\text{and} \quad -1 \leq r^* \leq 1$$

To verify this, we plotted r^* vs. $\min. \left(\beta, \frac{1}{c_{s_{\text{av}}}} \right) \cdot c_{a_{\text{av}}}$ and showed in fig. 5.2 that the relation is a line with slope of -1 that passes through the (0,0) point. Therefore, we conclude that the conjecture is acceptable.

Fig. 5.2 Correlation Coefficient that Minimizes the Relative Error ER as a Function of the Queuing System Parameters

5.4 Analysis of Results

5.4.1 Introduction

In this section, we check the accuracy of the diffusion approximation by comparing with analytical formulas and/or simulation runs. In sec. 5.4.2, we test the accuracy of the mean number of customers in an M/G/1 queueing system using analytical expressions. Then, we complement these results in sec. 5.4.3 by checking the accuracy of the distribution of GI/M/1 systems using simulation runs. Two values of the correlation coefficient (r) are used in this section, namely $r = -1$ and a value obtained from Eq. 5.24.

5.4.2 Accuracy of the Diffusion Approximation for M/G/1 Queueing Systems

The mean time in the system for an M/G/1 alternating priority with two classes of customers was derived by Avi-Itzhak et al [Av65], and Takacs [Ta68]. These means are given by

$$E[W_1] = \mu_1^{-1} + \frac{\lambda_1 E[S_1^2]}{2(1 - \rho_1)} + \frac{\lambda_2 E[S_2^2] (1 - \rho_1)^2 + \lambda_1 E[S_1^2] \rho_2^2}{2(1 - \rho_1)(1 - \rho_{\text{eff}})(1 - \rho_{\text{eff}} + 2\rho_1\rho_2)} \quad (5.25)$$

and

$$E[W_2] = \mu_2^{-1} + \frac{\lambda_2 E[S_2^2]}{2(1 - \rho_2)} + \frac{\lambda_1 E[S_1^2] (1 - \rho_2)^2 + \lambda_2 E[S_2^2] \rho_1^2}{2(1 - \rho_2)(1 - \rho_{\text{eff}})(1 - \rho_{\text{eff}} + 2\rho_1\rho_2)} \quad (5.26)$$

where ρ_i is the utilization of the server by class $i = \lambda_i / \mu_i$

$E[S_i^2]$ is the second moment of the service distribution of class i

$\rho_{\text{eff}} = \rho_1 + \rho_2 < 1$, i.e. system is stationary.

But from Little's formula [Li61]

$$E[n_i] = \lambda_i E[w_i]$$

We write $E[n_i]$ in terms of the parameters $\rho_1, \rho_2, \beta, c_1$ and c_2 as

$$E[n_1] = \rho_1 + \frac{\rho_1^2(c_1 + 1)}{2(1 - \rho_1)} + \frac{\frac{\rho_1 \rho_2}{\beta} (1 - \rho_1)^2 (c_2 + 1) + \rho_1^2 \rho_2^2 (c_1 + 1)}{2(1 - \rho_1)(1 - \rho_{\text{eff}})(1 - \rho_{\text{eff}} + 2\rho_1 \rho_2)} \quad (5.27)$$

$$E[n_2] = \rho_2 + \frac{\rho_2^2(c_2 + 1)}{2(1 - \rho_2)} + \frac{\frac{\rho_1 \rho_2}{\beta} (1 - \rho_2)^2 (c_1 + 1) + \rho_1^2 \rho_2^2 (c_2 + 1)}{2(1 - \rho_2)(1 - \rho_{\text{eff}})(1 - \rho_{\text{eff}} + 2\rho_1 \rho_2)} \quad (5.28)$$

where $\beta = \frac{\mu_2}{\mu_1}$

The mean number in the system is calculated from the diffusion approximation as

$$E[\bar{n}_i] = \hat{\rho}_i / (1 - \hat{\rho}_i) \quad (5.29)$$

where

$$\hat{\rho}_i = \exp \left\{ \frac{-2(1 - \rho_{\text{eff}})D_i}{\rho_{\text{eff}} c_{a_i} + c_{d_i}} \right\}, \quad i = 1, 2 \quad (5.30)$$

The relative error in the mean is

$$ER_1 = \frac{E[n_1] - E[\hat{n}_1]}{E[n_1]} \times 100 \quad (5.31)$$

We first discuss the case of $r = -1$. Figures 5.3 to 5.6 show the variation in the accuracy of the mean number of customers in the system, $(E(\hat{n}_1) + E(\hat{n}_2))$, as a function of α , β , ρ_{eff} , and the squared coefficient of variation of service time c_s (assuming $c_1 = c_2 = c_s$). We have the following observations:

- The relative error decreases as $\rho_{eff} \rightarrow 1$ (i.e. better accuracy for heavy traffic).
- The error is minimum for a system with $\alpha = .5$, $\beta = 1$, $c_a = c_s = 1$ and for this system $ER \rightarrow 0$ as $\rho_{eff} \rightarrow 1$.
- The accuracy is not sensitive to changes in α , especially for a system with $\beta = 1$ and $c_a = c_s = 1$ ($c_{a1} = c_{a2} = c_a$).
- The error is minimum for $\beta = 1$ and increases as we deviate from this value.
- We tend to underestimate the mean for systems with $c_s > 1$ (ER is positive). As for $c_s < 1$, the relative error can be positive or negative. The crossing point c_s^* (for $\rho_{eff} = 1$) depends essentially on the value of β (if $\beta > 1$ then $c_s^* < 1$, and vice-versa).

We proceed to study the relative error for the individual classes as a function of the system parameters. Figure 5.7 shows the effect of changing α on ER_1 for $\beta = 1$ and $c_a = c_s = 1$. We remark that the accuracy is not sensitive to α for low utilization ($\rho_{eff} = .8$). As we increase ρ_{eff} to .9,

Fig. 5.3 Accuracy vs. ρ_{eff} for $\beta = 1$ and $c_{a_1} = c_{a_2} = c_1 = c_2 = 1$

Fig. 5.4 Accuracy vs. β for $\alpha = 0.5$ and $c_{a_1} = c_{a_2} = c_1 = c_2 = 1$

Fig. 5.5 Accuracy vs. c_s for $\alpha = 0.5$, $\rho_{\text{eff}} = 0.9$ and $c_{a1} = c_{a2} = 1$.

Fig. 5.6 Accuracy vs. c_s for $\alpha = 0.5$, $\beta = 2$, and $c_{a1} = c_{a2} = 1$.

Fig. 5.7 ER_1 vs. α for $\beta = 1$, and $c_a = c_s = 1$

we find that the accuracy increases as $\alpha \rightarrow 1$. This is because the dominating class (large α) has a departure process that is closer to the "approximate" one, i.e. less error in c_{d1} . However, if ρ_{eff} is closer to 1 we find that error can become negative (overestimation), but it should go to 0 as $\alpha \rightarrow 1$ and $\rho_{\text{eff}} \rightarrow 1$ (dotted part).

Figure 5.8 to fig. 5.10 show the effect of changing β on the accuracy for different values of α and ρ_{eff} . From fig. 5.8, we find the error is large for small β and decreases as β increases. The reason for this is that a large β means more service time for class 1, therefore this class is dominating and is closer to the "approximate" output, i.e. large β has the same effect as large α . We notice also that $ER_1 \rightarrow 0$ as $\rho_{\text{eff}} \rightarrow 1$ for $\beta = 1$ and $\alpha = .5$. If $\alpha < .5$, the crossing point will be at $\beta > 1$ (see fig. 5.9). Figures 5.9 and 5.10 show the change of the accuracy of the diffusion estimates of the two classes as we vary β . We remark that the two sets of curves are opposite, i.e. ER_1 decreases while ER_2 increases as β takes larger values.

Figures 5.11 to 5.13 show the effect of varying c_s ($c_1 = c_2 = c_s$) on the relative error of the mean number in the system of class 1 customers. We find that:

- The error increases as c_s increases.
- We overestimate the mean for $c_s < 1$ and underestimate it for $c_s > 1$ (for a symmetrical system $\alpha = .5$, $\beta = 1$, $c_a = 1$).
- In general we overestimate for $c_s > c_s^*$ and underestimate for $c_s < c_s^*$, where $c_s^* < 1$ for $\alpha < .5$ or $\beta < 1$
 $c_s^* > 1$ for $\alpha > .5$ or $\beta > 1$
 (c_s^* corresponds to zero error at $\rho_{\text{eff}} = 1$).

Fig. 5.8 ER_1 vs. β for $\alpha = 0.5$, and $c_a = c_s = 1$

Fig. 5.9 ER_1 vs. β for $\alpha = 0.4$, and $c_a = c_s = 1$

Fig. 5.10 ER_2 vs. β for $\alpha = 0.4$ and $c_a = c_s = 1$

Fig. 5.11 ER_1 vs. c_s for $\alpha = 0.5$, $\beta = 1$, and $c_a = 1$

Fig. 5.12 ER_1 vs. c_s for $\alpha = 0.5$, and $c_a = 1$

Fig. 5.13 ER_1 vs. c_s for $\alpha = 0.4$, $\beta = 1$, and $c_a = 1$

In summary, the best accuracy is for a Symmetrical Markov System ($\alpha = .5$, $\beta = 1$, $c_a = c_s = 1$). The accuracy is quite satisfactory for large values of α and β , but c_s should be close to 1.

At this point, let us consider the effect of using r^* obtained from Eq. 5.24 on the accuracy of the diffusion approximation. Figures 5.14 to 5.16 show comparisons between the results for $r = -1$ and r^* . We find in general that using r^* resulted in a decrease in the positive error (underestimation) in the mean number of customers in the system (the total mean of both classes and the mean of class 1). Therefore, we conclude that the estimates with r^* are in general "better" than those with $r = -1$.

5.4.3 Accuracy of the Diffusion Approximation for GI/M/1 Queueing Systems

In this section, we compare the estimates obtained from the diffusion approximation with those obtained from simulation runs.

The first two cases are for M/M/1 queueing systems, one with equal service rates for both classes and the other with different service rates for each class. In Case 3 and Case 4 we changed the arrival process to hyperexponential and Erlang-2, respectively. We chose these cases because we will use GI/M/1 queueing systems in modelling channels in the following chapters. Furthermore, we studied the changes in the accuracy with the variation of the other parameters in the previous section.

Simulation

The simulation program is written in GPSS, and consists of approximately 67 blocks. Runs were performed for 20,000 transactions, and consumed 3.9 minutes of CPU time on the Carleton University XDS SIGMA-9 computer.

Fig. 5.14 ER vs. c_s for $\alpha = 0.5$, $\rho_{eff} = 0.9$, $c_{a1} = c_{a2} = 1$,
and Two Values for r .

Fig. 5.15 ER_1 vs. β for $\alpha = 0.5$, $c_a = c_s = 1$, and
Two Values for r

Fig. 5.16 ER_1 vs. c_s for $\alpha = 0.5$, $\rho_{eff} = 0.98$, $c_{a1} = c_{a2} = 1$ and Two Values for r .

Confidence interval for the mean is obtained using Crane and Iglehart's technique of independent cycles [CI74], and regeneration points start with the arrival of a customer to an empty system. The flowchart of the simulation program is presented in Appendix C.

Numerical Results

Table 5.1 and fig. 5.17 show the results obtained from simulation, diffusion approximation, and analytical formula, for an M/M/1 system with $\mu_1 = \mu_2$, $\alpha = .44$, and $\rho_{eff} = .887$. The accuracy of the diffusion is satisfactory here.

In case 2, we have different rates of service ($\beta = 2/3$) and we notice that the accuracy is better for class 2 (the class with higher service times). Table 5.2 and fig. 5.18 summarize the results for this case. Two sets of diffusion results are presented, one with correlation coefficient (r) equal -1, the second with r^* calculated from Eq. 5.24.

Similarly, table 5.3 and fig. 5.19 summarize the results for the case with hyperexponential arrival times, while table 5.4 and fig. 5.20 give the results for the Erlang-2 arrival case.

Poisson arrival

 $\alpha = .44$ Exponential service times ($\mu_1^{-1} = \mu_2^{-1} = 20$) $\beta = 1$ $\rho_{\text{eff}} = .89$ $c_{d1} = c_{d2} = .51$

		Queue 1	Queue 2	Total
<u>Sim</u>	mean	3.66	3.87	7.53
	90% Conf. Int. for mean	[3.0-4.21]	[3.31-4.33]	
	var	24.5	20.5	
<u>Diff</u>	$\hat{\rho}_i$	.77	.78	
$r = -1$	mean	3.33	3.56	6.89
	ER_i	9	8	8.5
	var	14.5	16.2	
	ERV_i	40.8	21	
<u>Analytical</u>	mean	3.83	4.02	7.85

Table 5.1 Accuracy of the Diffusion Approximation for Case 1

Fig. 5.17 Marginal Queue Length Distributions for Case 1

Poisson arrival

 $\alpha = .44$ Exponential service times ($\mu_1^{-1} = 20, \mu_2^{-1} = 30$) $\beta = 2/3$ $\rho_{\text{eff}} = .9$ $c_{d1} = c_{d2} = .55$

		Queue 1	Queue 2	Total
<u>Sim</u>	mean	5.94	4.46	10.40
	90% Conf. Int. for mean	[4.26-7.26]	[3.51-5.17]	
	var	75.7	30.3	
<u>Analytical</u>	mean	5.47	4.22	9.69
<u>Diff</u>	$\hat{\rho}_i$	.792	.804	
$r = -1$	mean	3.82	4.71	7.93
	ER_i	35.7	7.8	23.75
	var	18.2	20.6	
	ERV_i	75.9	32.7	
<u>Diff</u>	$\hat{\rho}_i$	.823	.831	
$r = -.7$	mean	4.64	4.91	9.55
	ER_i	21.9	-10.1	8.2
	var	26.2	29.1	
	ERV_i	65.4	4	

Table 5.2 Accuracy of the Diffusion Approximation for Case 2

Fig. 5.18 Marginal Queue Length Distributions for Case 2

Hyperexponential arrival times ($c_{a_1} = c_{a_2} = 2$) $\alpha = .44$
 Exponential service times ($\mu_1^{-1} = \mu_2^{-1} = 20$) $\beta = 1$

$$\rho_{\text{eff}} = .881 \quad c_{d_1} = c_{d_2} = .51$$

		Queue 1	Queue 2	Total
<u>Sim</u>	mean	5.96	6.69	12.65
	90% Conf. Int. for mean	[4.51-7.06]	[4.77-8.19]	
	var	73.96	92.16	
<u>Analytical</u>		3.6	3.8	7.4
<u>Diff</u>	$\hat{\rho}_i$	.871	.876	
$r = -1$	mean	6.76	7.06	13.8
	ER_i	-13.4	-5.5	-9.1
	var	52.4	56.5	
	ERV_i	29.1	38.7	

Table 5.3 Accuracy of the Diffusion Approximation for Case 3

Fig. 5.19 Marginal Queue Length Distributions for Case 3

Erlang-2 arrival times ($c_{a1} = c_{a2} = .5$) $\alpha = .44$ Exponential service times ($\mu_1^{-1} = \mu_2^{-1} = 20$) $\beta = 1$ $\rho_{\text{eff}} = .881$ $c_{d1} = c_{d2} = .51$

		Queue 1	Queue 2	Total
<u>Sim</u>	mean	3.18	3.48	6.66
	90% Conf. Int. for mean	[2.53-3.73]	[2.81-4.04]	
	var	21.9	21.6	
<u>Analytical</u>		3.6	3.8	7.4
<u>Diff</u>	$\hat{\rho}_1$	.57	.594	
$r = -1$	mean	1.32	1.47	3.2
	ER_1	58.5	57.8	52.
	var	3.07	3.62	
	ERV_1	85.8	83.2	
<u>Diff</u>	$\hat{\rho}_1$	.703	.715	
$r = -.5$	mean	2.37	2.51	4.88
	ER_1	25.5	27.9	26.7
	var	8.1	8.7	
	ERV_1	63.	59.7	

Table 5.4 Accuracy of the Diffusion Approximation for Case 4

Fig. 5.20 Marginal Queue Length Distributions for Case 4

Accuracy vs. c_{aav}

The results of cases 1, 3 and 4 are plotted in fig. 5.21 to show the change in the accuracy with the variation of c_a . From this figure we find that the accuracy is better for $c_a \geq 1$. For $c_a < 1$, the error is positive (underestimation) and high. This behavior is similar to the case of FCFS queueing systems (fig. 3.21), therefore we proposed to modify r similar to the FCFS case, i.e. r^* proportional to c_{aav} , as has been shown in Eq. 5.24.

5.5 Concluding Remarks

In this chapter, we derived a diffusion approximation for the distribution of the number of customers of each class in a GI/G/1 queueing system with 2 classes and alternating priority. We checked the accuracy of the approximation against analytical formulas and simulation runs for different values of the system parameters (α , β , c_s , c_a and ρ_{eff}). We found that the best accuracy is for the Symmetrical Markov System (same arrival and service rates); also that the accuracy is quite satisfactory for the class with the highest probability of arrival or the largest mean for the service time. To improve the accuracy for other values of β , c_a , and c_s we proposed a conjecture to calculate the correlation coefficient to be used in the diffusion equations.

Fig. 5.21 Accuracy vs. c_a for $\alpha = 0.44$, $\beta = 1$, and $c_s = 1$

6. NETWORKS OF QUEUES

6.1 Introduction

In the preceding three chapters, we developed diffusion approximations for the distributions of the number of customers in GI/G/1 queueing systems with two classes of customers. The server has one of the following service disciplines: (1) Preemptive-Resume, (2) FCFS, and (3) Alternating Priority. In this chapter, we develop two approximate algorithms for the solution of open and closed networks with two classes of customers and the previously mentioned service disciplines.

A review of the literature on the exact analytical solutions for open networks is presented in sec. 6.2.1. Then, approximate solutions are discussed in sec. 6.2.2 with a description of Kobayashi's technique [Ko74, RK74]. In sec. 6.3, Reiser and Kobayashi's algorithm for open networks [RK74] is extended to networks with two classes of customers and the service disciplines described in Chapters 3, 4, and 5. Numerical results of the algorithm and discussions of their accuracy are presented in sec. 6.4. In sec. 6.5.1, we review the literature on analytical and approximate solutions for closed networks. Then in sec. 6.5.2 we present an algorithm to identify bottleneck(s) in a closed network and transform the "unsaturated" part of the network to an open one. We use this algorithm for two numerical examples in sec. 6.6.

6.2 Review of Literature

6.2.1 Exact Solutions for Open Networks

Recently, computer systems' researchers showed interest in queueing networks and the techniques used for analysing them because of their

importance in modelling complex computer configurations with multi-programming and/or time sharing environment. The pioneer research in this area is due to R.R.P. Jackson who studied a network of two queueing centers in tandem, an M/M/1 feeding a single exponential server [JR54], and showed that the equilibrium joint distribution of the number of customers in each center is the product of two marginal distributions each of an M/M/1 queueing system. This indicated that the output from the first center must be Poisson if the arrival to that center is Poisson and the service times are exponential. This result was proved later by Burke [Bu56] for the M/M/S queueing systems.

J.R. Jackson studied an arbitrary Markovian network of M/M/S queueing centers with external arrivals and departures (referred to as an open network). External arrivals are Poisson, and transitions between service centers are Markovian and ergodic. Jackson [JJ57] showed that each center behaves as if its input were Poisson, and the joint equilibrium distribution of the number of customers in each center is the product of the marginal distributions for each center. This is known as "Jackson Theorem". Later, Jackson showed that the joint equilibrium distribution still has a product form if the arrival rates are allowed to depend on the total number of customers in the network and the service rate at any center is permitted to depend arbitrarily on the number of customers presently in this center [JJ63].

Gordon and Newell [GN67] modelled an arbitrary Markovian network where a fixed number of customers pass repeatedly through the service centers of the network (referred to as closed network), and showed that the joint

equilibrium distribution has a product form. However, there is a normalizing constant that is a function of all possible states of the network. Buzen [Bu73] proposed an iterative algorithm for the calculation of this normalizing constant. Moore [Mo71] used a closed Markovian network to model the Michigan Terminal System (MTS), and verified his results by measurements. The predicted mean response time was within 10% of the corresponding measured values. Scherr [Sc67] used the Markovian models for the MIT CTSS time-sharing system.

Chandy [Ch72] developed the "local balance" concept and showed that it could be used to obtain product form solutions for a large class of queueing networks. Baskett et al [Ba75] generalized the previous results to the case of networks of queues with multi-class customers. They derived the joint equilibrium state probabilities for a network containing N service centers, and R classes of customers. Each customer belongs to a single class while at a service center, but may change class and service centers according to a fixed transition probability. The service centers have one of the following disciplines:

1. FCFS; all customers have the same exponential service time distribution.
2. Processor Sharing (PS); the service time distributions have rational transforms and different rates for each class.
3. Infinite Server (IS); the number of servers is equal to the number of customers at any time, and the service time distributions are similar to those described in 2.
4. Last Come First Served (LCFS); Preemptive-Resume, and the service time distributions are similar to those described in 2.

For these service disciplines Muntz [Mu73] showed that the departure process is Poisson provided the arrival is also Poisson. The joint equilibrium distribution for a network that satisfies the previous conditions has a product form [Ba75].

Chandy, Herzog and Woo [Ch75a] developed the "Norton's Theorem" technique for solving locally balanced networks by "short-circuiting" the queueing center we are interested in, and calculating the parameters of the "composite queue" that represents the remainder of the network. Chandy [Ch77] introduced the concept of "Station balance" which is a sufficient condition for "local balance". They showed that to have the product form solution, all service centers must begin to serve new customers immediately on their arrival (e.g. LCFS P.R., IS, P.S.), or have service distributions which are negative exponential with the same mean service rate for all classes.

6.2.2 Approximate Solutions for Open Networks

For general networks with queueing centers which do not satisfy the previous conditions, it has not been proven that the product form solutions hold. Simulation techniques are used for the general network (e.g. [CM75, Mc75, CF75]), and approximate techniques have been developed for some classes of networks. Chandy et al [Ch75b] presented an iterative approximate technique for solving networks with non-exponential service distributions which may depend on customer class. Another iterative technique for general networks is due to Wallace and Rosenberg [WR66]. These techniques are relatively expensive to use because of their iterative nature. Therefore, it is not practical to use them in designing networks that involve many parameters. Inexpensive approximate techniques were

developed by Sauer and Chandy [Sa75, SC75] for central server models. These techniques use transformation of the original problem to a simpler network with 2 queues which has a product form solution.

A different approximate approach is based on diffusion approximation and was developed by Kobayashi for open and closed networks under heavy traffic with one class of customers and FCFS queueing centers [Ko74]. His technique uses the multidimensional diffusion equation and a set of reflecting boundary conditions. We consider this technique in some detail in the remainder of this section.

Kobayashi's Approach for Open Networks

Consider a network with N FCFS service centers, in which service times at server i are i.i.d. with mean $1/\mu_i$ and variance σ_i^2 , $i = 1, 2, \dots, N$. We introduce two more service centers to the network, namely a source center marked 0 and a sink center marked $N+1$. The interarrival times of external customers, generated in the source center, are i.i.d. with mean $1/\lambda = 1/\mu_0 = 1/\lambda_0$ and variance σ_0^2 . After receiving service at center i , customers are transferred^a instantaneously to server j with transition probability r_{ij} , where

$$\sum_{j=1}^N r_{0j} = 1 \quad \text{and} \quad \sum_{j=1}^{N+1} r_{ij} = 1, \quad j \leq i \leq N \quad (6.1)$$

The transition probabilities in the network are Markovian, i.e., they may depend on the present state and not on the past history.

Let $A_i(t)$ and $D_i(t)$ represent the cumulative numbers of arrivals and departures at center i up to time t , and let $D_{ji}(t)$ represent the number of customers that joined center i after receiving service from center j ,

then

$$A_i(t) = \sum_{j=0}^N D_{ji}(t) \quad (6.2)$$

$$D_i(t) = \sum_{k=1}^{N+1} D_{ik}(t) \quad (6.3)$$

where $1 \leq i \leq N$

Following arguments similar to those in Chapter 2, the change in queue length i in time dt is given by

$$\Delta Q_i(t) = \Delta A_i(t) - \Delta D_i(t) \quad (6.4)$$

and is normally distributed with the mean and the covariance given by

$$E[\Delta Q_i(t)] = b_i dt \quad (6.5)$$

$$\text{cov}[\Delta Q_i(t)] = a_{ij} dt, \quad 1 \leq i, j \leq N \quad (6.6)$$

Kobayashi [Ko74] derived the following expressions for b_i and a_{ij}

$$b_i = \sum_{k=0}^N r_{ki} \mu_k - \mu_i \quad (6.7)$$

and

$$a_{ij} = \sum_{k=0}^N ((c_k - 1) \mu_k) r_{ki} r_{kj} +$$

$$\{c_i \mu_i + \sum_{k=0}^N r_{ki} \mu_k\} \delta_{ij} - c_i \mu_i r_{ij} - c_j \mu_j r_{ji} \quad (6.8)$$

where $c_i = \sigma_i^2 \mu_i^2$

and δ_{ij} is the Kronecker delta.

The N-dimensional discrete-state process, $Q(t) = \{Q_i(t)\}$, can then be approximated by an N-dimensional continuous process $\bar{X}(t)$ which satisfies the stochastic differential equation (see Eq. 2.17).

$$d\bar{X}(t) = b dt + (a dt)^{1/2} dW(t) \quad (6.9)$$

where $W(t)$ is an N-dimensional Wiener Process with zero mean and unit covariance matrix,

b is an N-dimensional column vector,

and a is a positive definite matrix.

Given that $\bar{x}(t)$ is equal to $\bar{x}_0$ at the initial time 0, then the solution of the stochastic differential equation with no boundary conditions would lead to the following conditional probability density at time t

$$p(\bar{x}, t) = (2\pi \det|at|)^{-1/2} \exp\left\{-\frac{1}{2}t (\bar{x} - \bar{x}_0 - bt)^T a^{-1} (\bar{x} - \bar{x}_0 - bt)\right\} \quad (6.10)$$

which satisfies the multidimensional diffusion equation (Eq. 2.18) which is written as

$$\frac{\partial}{\partial t} p(\bar{x}, t) = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N a_{ij} \frac{\partial^2}{\partial x_i \partial x_j} p(\bar{x}, t) + \sum_{i=1}^N b_i \frac{\partial}{\partial x_i} p(\bar{x}, t) \quad (6.11)$$

By setting the left hand side of Eq. 6.11 to zero, and introducing N reflecting boundaries at $x_i = 0$, $i = 1, 2, \dots, N$, it can be shown that the steady-state solution for the joint distribution is given by

$$p(\bar{x}) = \prod_{i=1}^N |\gamma_i| \exp(\gamma_i x_i) \quad (6.12)$$

where $\gamma^T = 2 \alpha^{-1} \beta$ is a vector with negative elements. The last expression has the product form. By discretizing it, we obtain the approximate solution for the joint distribution of the number in N centers as

$$\hat{p}(n_1, n_2, \dots, n_N) = \prod_{i=1}^N \hat{p}_i(n_i) \quad (6.13)$$

$$\text{where} \quad \hat{p}_i(n_i) = (1 - \hat{\rho}_i) \hat{\rho}_i^{n_i} \quad (6.14)$$

$$\text{and} \quad \hat{\rho}_i = \exp \gamma_i \quad (6.15)$$

Kobayashi [Ko74] improved his previous solution by obtaining expressions for the probability of an idle server $p_i(0)$. These are obtained by solving the following set of simultaneous linear equations.

$$e_i = \sum_{k=0}^N e_k r_{ki} \quad 1 \leq i \leq N \quad (6.16)$$

where e_i is the average number of visits that a customer makes to the i^{th} server before leaving the network,

and $e_0 = 1$

$$\text{Then, } p_i(0) = 1 - e_i \frac{\mu_0}{\mu_i} = 1 - \rho_i \quad (6.17)$$

where ρ_i represents the utilization of server i .

Therefore

$$\hat{p}_i(n) = \begin{cases} 1 - \rho_i & , n = 0 \\ \rho_i (1 - \hat{\rho}_i) \hat{\rho}_i^{n-1} & , n \geq 1 \end{cases} \quad (6.18)$$

The previous approximation was tried for a number of cases and yielded satisfactory results [Ko74].

The product form solution given by Eq. 6.12 motivated Reiser and Kobayashi [RK74] to develop a simplified algorithm that solves for the marginal distribution of the number in each center first, and then obtains the joint distribution from the product form. This algorithm requires considerably less computation, with some sacrifice in the accuracy.

We summarize Reiser and Kobayashi's algorithm for open networks in the following:

1. Calculate the arrival rate to center i , and the utilization of this center as follows:

a. Solve the set of linear equations

$$e_i = r_{oi} + \sum_{j=1}^N e_j r_{ji} \quad (6.19)$$

for $e_i \quad 1 \leq i \leq N$

b. The arrival rate to center i is given by

$$\lambda_i = \mu_0 e_i \quad (6.20)$$

where μ_0 is the total arrival rate from the outside source

c. The utilization is

$$\rho_i = \lambda_i / \mu_i \quad (6.21)$$

2. Calculate the squared coefficient of variation of the interarrival times to center i as

$$c_a^{(i)} = 1 + \sum_{j=0}^N (c_j - 1) \cdot r_{ji}^2 \lambda_j / \lambda_i \quad (6.22)$$

3. Calculate the marginal distribution of the number of customers in center i using Eqs. 2.33 and 2.32 as follows

$$\hat{p}_i(n) = \begin{cases} 1 - \rho_i & , n = 0 \\ \rho_i (1 - \hat{\rho}_i) \hat{\rho}_i^{n-1} & , n \geq 1 \end{cases} \quad (6.23)$$

where

$$\hat{\rho}_i = \exp \left(\frac{-2(\mu_i - \lambda_i)}{c_a^{(i)} \lambda_i + c_i \mu_i} \right)$$

Gelenbe and Pujolle [GP76] improved the previous algorithm by replacing the coefficients c_j in Eq. 6.22 by the squared coefficient of variation of the interdeparture times at that center. They proposed an algorithm to calculate the parameters of the departure processes.

6.3 Open Networks with More Than One Class of Customers

6.3.1 Introduction

In this section, we extend Reiser and Kobayashi's approach, described in the previous section, to the case of open networks with two classes of customers and the service disciplines described in Chapters 3, 4, and 5.

The network under consideration has a finite number N of service centers. Each center has one or two classes of customers, a general service time distribution for each class, and the service discipline could be FCFS, Preemptive-Resume, or alternating priority. Therefore, a customer in the network is identified with the pair (i, l) , where the first label identifies the service center, and the second one identifies the class of the customer. A customer will retain his class as long as he stays in the center but may change his class after receiving service and join the same center or another center. The transition probability $r_{i,l;j,m}$ is the probability that a customer of class l at center i will join class m customers in center j upon receiving service ($1 \leq i, j \leq N$, and $l, m = 1$ or 2).

The transition matrix $R = \{r_{i,l;j,m}\}$ defines a Markov chain where the states are identified with the pair (i, l) . For simplicity of presentation, we assume the chain is irreducible, i.e. every state can be reached from every other state. If not, we decompose the network to sub-chains which are irreducible, and solve each sub-chain separately.

External arrivals are generated from a renewal process source at a mean rate λ and squared coefficient of variation for interarrival times (c_0). Class l at center i receives external customers with probability $q_{i,l} = \frac{\lambda}{r_{0,1;i,l}}$ (We assumed that there is only one class of customers

generated from the source center, i.e., $r_{0,2;i,l} = 0$ for $1 \leq i \leq N$, $l = 1, 2$). Let $\lambda_{i,l}$ denote the mean arrival rate of class l customers at center i , and let $\mu_{i,l}$ be the mean service rate of those customers.

$$\text{If } \rho_i = \sum_{l=1}^2 \frac{\lambda_{i,l}}{\mu_{i,l}} < 1 \quad \text{for } 1 \leq i \leq N \quad (6.24)$$

then the Markov chain is ergodic and a steady-state solution exists.

Therefore, the set of linear equations

$$\lambda_{i,l} = \Lambda q_{i,l} + \sum_{j=1}^N \sum_{m=1}^2 \lambda_{j,m} r_{j,m;i,l} \quad (6.25)$$

has a unique solution.

A parallel set of equations can be obtained by dividing Eq. 6.25 by Λ and assuming

$$e_{i,l} = \lambda_{i,l} / \Lambda \quad (6.26)$$

then,

$$e_{i,l} = q_{i,l} + \sum_{j=1}^N \sum_{m=1}^2 e_{j,m} r_{j,m;i,l} \quad (6.27)$$

The variable $e_{i,l}$ is interpreted as the average number of visits to class l at center i by a customer during his stay in the network.

6.3.2 The Arrival Processes

The arrival process of class l customers to center i is the sum of the departure processes from any class m in any center j that have non-zero transition probabilities $r_{j,m;i,l}$.

Let the mean and the squared coefficient of variation of the interdeparture times of class m customers at center j be $1/\mu_d^{j,m}$ and $c_d^{j,m}$, respectively, and $F_Y(y)$ and $\phi_Y(\omega)$ are the PDF and the characteristic function of the r.v. Y which represents the interdeparture times. Assume $A_{j,m;i,l}(t)$ is the cumulative number of arrivals to class l in center i , from customers leaving class m in center j , and let $X_{j,m;i,l}$ be a r.v. that represents the interevent times of the process $A_{j,m;i,l}(t)$, then,

$$\begin{aligned} P[X_{j,m;i,l} \leq x] &= P[Y^1 \leq x \text{ and the first departure from } (j,m) \\ &\quad \text{goes to } (i,l)] \\ &\quad + P[Y^1 + Y^2 \leq x \text{ and the first departure from} \\ &\quad (j,m) \text{ does not go to } (i,l), \text{ but the second does}] \\ &\quad + \dots \end{aligned} \quad (6.28)$$

where $(Y^1, Y^2, Y^3, \dots)$ are independent r.v.'s with the same PDF of the interdeparture times of class m customers at center i .

Thus,

$$P[X_{j,m;i,l} \leq x] = r_{j,m;i,l} \sum_{n=0}^{\infty} (1 - r_{j,m;i,l})^n \times \{F_Y(x)\}^{n+1} \quad (6.29)$$

where $\{F_Y(x)\}^{n+1}$ is an $(n+1)$ -fold convolution of $F_Y(x)$.

The characteristic function of $X_{j,m;i,l}$ is $\phi_X(\omega)$, and is given as

3

OF/DE

3

$$\phi_X(\omega) = \frac{r_{j,m,i,l} \phi_Y(\omega)}{1 - (1 - r_{j,m,i,l}) \phi_Y(\omega)} \quad (6.30)$$

Using the definition of the n^{th} moment given in Eq. 4.7, we obtain the first and second moments as

$$\overline{X}_{j,m,i,l} = \frac{1}{r_{j,m,i,l} \mu_d^{j,m}} \quad (6.31)$$

$$\text{and } \overline{X_{j,m,i,l}^2} = \frac{1}{r_{j,m,i,l}^2} \{ r_{j,m,i,l} \overline{Y^2} + 2(1 - \alpha) / (\mu_d^{j,m})^2 \} \quad (6.32)$$

where $\overline{Y^2}$ is the second moment of the d.v. Y . Therefore, the squared coefficient of variation of $X_{j,m,i,l}$ is

$$c_d^{j,m,i,l} = 1 + r_{j,m,i,l} (c_d^{j,m} - 1) \quad (6.33)$$

From the central limit theorem of renewal processes, we find the following mean and variance for the number of arrivals to (i,l) from (j,m) in time dt

$$E[\Delta D_{j,m,i,l}(t)] = r_{j,m,i,l} \mu_d^{j,m} dt \quad (6.34)$$

$$\text{var}[\Delta D_{j,m,i,l}(t)] = c_d^{j,m,i,l} r_{j,m,i,l} \mu_d^{j,m} dt \quad (6.35)$$

where $\Delta D_{j,m,i,l}(t)$ is the number of customers that join class l at center i after leaving class m at center j in time dt .

The corresponding infinitesimal mean and variance can be obtained using definitions given in Eqs. 2.5 and 2.6. The total arrival to (i,l) is composed of the sum of all the customers that leave (j,m) ($0 \leq j \leq N$ and $m = 1, 2$) and go to (i,l) . Under heavy traffic, the arrivals from different queues (j,m) are independent, and we obtain the mean and variance of the total arrival using expressions given by Eqs. 6.34 and 6.35 weighted by the utilization ρ_j of center j

$$\lambda_{i,l} = \sum_{j=0}^N \sum_{m=1}^2 \rho_j r_{j,m;i,l} \mu_d^{j,m} \quad (6.36)$$

where $\rho_0 = 1$,

$$\begin{aligned} \text{and} \quad \frac{\text{var}[\Delta A_{i,l}(t)]}{dt} &= \sum_{j=0}^N \sum_{m=1}^2 \text{var}[\Delta D_{j,m;i,l}(t)] \rho_j \\ &= \sum_{j=0}^N \sum_{m=1}^2 \rho_j c_d^{j,m;i,l} r_{j,m;i,l} \mu_d^{j,m} \\ &= \sum_{j=0}^N \sum_{m=1}^2 [(c_d^{j,m} - 1) r_{j,m;i,l} + 1] \end{aligned}$$

$$r_{j,m;i,l} \mu_d^{j,m} \rho_j \quad (6.37)$$

where $\lambda_{i,l}$ is the total mean arrival rate of class l customers to center i

and $\Delta A_{i,l}(t)$ is the total number of customers that join class l at center i in time dt .

Therefore, the squared coefficient of variation of the interarrival time is expressed as

$$\begin{aligned}
 c_a^{i,l} &= \frac{\text{var}[\Delta A_{i,l}]/dt}{\lambda_{i,l}} \\
 &= \frac{1}{\lambda_{i,l}} \sum_{j=0}^N \sum_{m=1}^2 [(c_d^{j,m} - 1) r_{j,m;i,l} + 1] \\
 &\quad r_{j,m;i,l} \nu_d^{j,m} \rho_j \\
 &= 1 + \sum_{j=0}^N \sum_{m=1}^2 (c_d^{j,m} - 1) r_{j,m;i,l}^2 \lambda_{j,m} / \lambda_{i,l} \quad (6.38)
 \end{aligned}$$

6.3.3 The Algorithm

The following is the algorithm for the solution of open networks with two classes of customers. The algorithm includes equations derived in Chapters 3, 4, 5 and 6.

1. Solve the set of linear equations

$$\lambda_{i,l} = \Lambda q_{i,l} + \sum_{j=1}^N \sum_{m=1}^2 \lambda_{j,m} r_{j,m;i,l} \quad (6.25)$$

for values of $\lambda_{i,l}$

$$\text{and calculate } \rho_i = \sum_{l=1}^2 \lambda_{i,l} / \mu_{i,l} \quad (6.39)$$

2. Calculate the squared coefficient of variation for the inter-departure times of class l at center i ($c_d^{i,l}$) as follows:

- For the high priority customers in Preemptive-Resume queueing centers and for the customers in FCFS single class queueing centers

$$c_d^{i,l} = c_{i,l}$$

where $c_{i,l}$ is the squared coefficient of variation for the service times of class l at center i .

- For both classes in FCFS queueing centers and for the low priority class in Preemptive-Resume queueing centers

$$c_d^{i,l} = \frac{\alpha_i^2 c_{i,l} + (\alpha_i c_{i,m} + 1)(1 - \alpha_i)/\beta_i^2}{(\alpha_i + \frac{1 - \alpha_i}{\beta_i})^2}, \quad m \neq l \quad (4.11)$$

$$\text{where } \alpha_i = \frac{\lambda_{i,l}}{\lambda_{i,l} + \lambda_{i,m}}$$

$$\beta_i = \mu_{i,m} / \mu_{i,l}$$

- For both classes in alternating priority queueing centers

$$c_d^{i,l} = c_d^{i,m} = \frac{\alpha_i^2 c_{i,l} + (\frac{1 - \alpha_i}{\beta_i})^2 c_{i,m}}{(\alpha_i + \frac{1 - \alpha_i}{\beta_i})^2}, \quad m \neq l \quad (5.9)$$

3. Calculate the squared coefficient of variation for the inter-arrival times of class l at center i as

$$c_a^{i,l} = 1 + \sum_{j=0}^N \sum_{m=1}^2 (c_d^{j,m} - 1) r_{j,m;i,l}^2 \frac{\lambda_{j,m}}{\lambda_{i,l}} \quad (6.38)$$

where $j = 0$ indicates the external arrivals.

4. The diffusion approximation of the distribution of class customers at center i is given by

$$\hat{p}(n_{i,l}) = (1 - \hat{\rho}_{i,l}) \hat{\rho}_{i,l}^{n_{i,l}}, \quad i = 1, 2, \dots, N$$

$$l = 1, 2 \quad (4.44)$$

where $\hat{\rho}_{i,l} = \exp \gamma_{i,l}$ (4.45)

$$\gamma_{i,l} = \frac{-2(1 - \rho_i^*)}{\rho_i^* c_a^{i,l} + c_d^{i,l}} D_{i,l} \quad (4.50)$$

and ρ_i^* is the total utilization of center i (ρ_i), except for the high priority class (l) in Preemptive-Resume queueing centers where

$$\rho_i^* = \rho_{i,l} = \lambda_{i,l} / \mu_{i,l}$$

The value of $D_{i,l}$ is calculated as follows:

- For the both classes in Preemptive-Resume queueing centers and for one class FCFS centers

$$D_{i,l} = 1$$

- For two class FCFS or alternating priority centers

$$D_{i,l} = \frac{c_a^{i,m} \rho_i^* + c_d^{i,m} - r \sqrt{c_d^{i,m} c_d^{i,l} \frac{1 - \alpha_i}{\alpha_i}}}{c_a^{i,m} \rho_i^* + c_d^{i,m} - \frac{r^2 c_d^{i,m} c_d^{i,l}}{c_a^{i,l} \rho_i^* + c_d^{i,l}}}, \quad m \neq l \quad (4.52)$$

and $r = -1$; other values for r to improve the accuracy of the approximation are obtained empirically (see Eq. 4.60 for FCFS and Eq. 5.24 for alternating priority).

5. The mean and variance of the number of class l customers in center i are obtained as

$$\bar{Q}_{i,l} = E[\hat{n}_{i,l}] = \hat{\rho}_{i,l} / (1 - \hat{\rho}_{i,l})$$

$$\text{var}[\hat{n}_{i,l}] = \hat{\rho}_{i,l} / (1 - \hat{\rho}_{i,l})^2$$

6. The mean time spent in the network between queue (i,l) and queue (j,m) is calculated as

$$T_{i,l,j,m} = \sum_k \sum_n \frac{\bar{Q}_{k,n}}{\lambda_{k,n}} = \frac{1}{\Lambda} \sum_k \sum_n \frac{\bar{Q}_{k,n}}{a_{k,n}} \quad (6.40)$$

(for all (k,n) pairs between (i,l) and (j,m))

The extension of this algorithm to networks with more than two classes of customers is straightforward.

6.4 Numerical Results for Open Networks

6.4.1 Introduction

In this section, we present two examples of open networks to show the use of the previous algorithm. All the results are compared to simulation runs, and confidence intervals are obtained using Crane and Iglehart's technique of independent cycles [CI74]. For open networks, cycles are assumed to start with the arrival of a customer to an empty network (all servers are idle).

6.4.2 Example 1

Consider the open network shown in fig. 6.1 which consists of 3 service centers. Each center has two classes of customers marked 1 and 2. Servers 1 and 3 have Preemptive-Resume service discipline, while server 2 deals with all customers on an FCFS basis. This network can be a model for two processors sharing a disk. The processors are modelled by Preemptive-Resume, and the disk is modelled by an FCFS server attending I/O requests from the two processors. The states of the network are given by the pair (i, l) where the first number identifies the center and the second identifies the class. Class 1 in the Preemptive-Resume centers has the higher priority. Service time distributions are all negative exponential with the following means for centers 1 and 3:

$$\begin{array}{ll} \mu_{1,1}^{-1} = 4 & \mu_{1,2}^{-1} = 32 \\ \mu_{3,1}^{-1} = 2 & \mu_{3,2}^{-1} = 20 \end{array}$$

As for center 2, we assume $\mu_{2,1}$ is a variable and $\mu_{2,2} = \mu_{2,1}$.

Fig. 6.1 Open Network of Example 1

External arrivals are Poisson, and R is assumed equal to 1, therefore

$\bar{q} = \{q_{i,l}\}$ is given by

$$\bar{q} = \{.5, 0, 0, 0, .5, 0\}$$

The transition matrix $\{r_{i,l;j,m}\}$ is (8 X 8) matrix (including nodes for source (o) and sink (s)), and is given by

$i,l \backslash j,m$	o	1,1	1,2	2,1	2,2	3,1	3,2	s
o	o	.5	o	o	o	.5	o	o
1,1	o	o	.5	o	.5	o	o	o
1,2	o	o	o	o	o	.5	o	.5
2,1	o	o	o	o	o	o	1	o
2,2	o	o	1	o	o	o	o	o
3,1	o	o	o	.5	o	o	.5	o
3,2	o	.5	o	o	o	o	o	.5
s	o	o	o	o	o	o	o	o

We proceed to apply the algorithm for the case where $\Lambda^{-1} = 40$ and

$$\mu_{2,1}^{-1} = 36$$

Step 1

From Eq. 6.27 we find the following set of simultaneous equations:

$$e_{1,1} = .5 + .5 e_{3,2}$$

$$e_{1,2} = .5 e_{1,1} + e_{2,2}$$

$$e_{2,1} = .5 e_{3,1}$$

$$e_{2,2} = .5 e_{1,1}$$

$$e_{3,1} = .5 + .5 e_{1,2}$$

$$e_{3,2} = e_{2,1} + .5 e_{3,1}$$

The solution is:

$$e_{1,1} = e_{1,2} = e_{3,1} = e_{3,2} = 1$$

$$e_{2,1} = e_{2,2} = .5$$

Thus, from Eq. 6.26 we find the mean arrival rates to the different queues:

$$\lambda_{1,1} = \lambda_{1,2} = \lambda_{3,1} = \lambda_{3,2} = \Lambda = 1/40$$

$$\lambda_{2,1} = \lambda_{2,2} = .5\Lambda = 1/80$$

The utilizations of the various servers are:

$$\rho_1 = 4\Lambda + 32\Lambda = .9$$

$$\rho_2 = 36\Lambda/2 + 36\Lambda/2 = .9$$

$$\rho_3 = 2\Lambda + 20\Lambda = .55$$

Step 2

The squared coefficients of variation of the interdeparture times are calculated as follows:

- For centers 1 and 3, we use Eq. 3.17 for the low priority class.

Given $\beta_1 = 0.125$

$$\text{therefore, } c_d^{1,1} = 1$$

$$c_d^{1,2} = .827$$

$$\text{similarly, } \beta_3 = 0.1$$

$$\text{and } c_d^{3,1} = 1$$

$$c_d^{3,2} = .851$$

- For center 2, we use Eqs. 4.11 and 4.14

Since, $\beta_2 = 1$

then, $c_d^{2,1} = c_d^{2,2} = 1$

Step 3

Using Eq. 6.38, we calculate the squared coefficients of variation for the arrival processes:

$$c_a^{1,1} = 1 + (.851 - 1) \times (.5)^2 = .963$$

$$c_a^{1,2} = c_a^{2,1} = c_a^{2,2} = c_a^{3,2} = 1$$

$$c_a^{3,1} = 1 + (.827 - 1)(.5)^2 = .957$$

Step 4

The values of $\hat{\rho}_{i,l}$ are calculated as

i, l	1,1	1,2	2,1	2,2	3,1	3,2
$\hat{\rho}_{i,l}$	.194	.891	.801	.801	.163	.526

Step 5

The mean number of customers of class l in center i ($\bar{Q}_{i,l}$) is calculated as

i, l	1,1	1,2	2,1	2,2	3,1	3,2
$\bar{Q}_{i,l}$	.124	8.15	4.02	4.02	.06	1.11

Step 6

The mean time spent between (1,1) and (3,2) is calculated using Eq. 6.40 as follows:

$$T_{1,1;3,2} = 40 \left\{ .124 + 8.15 + \frac{4.02}{.5} + .06 + 1.11 \right\}$$

$$= 699.36$$

Similarly,

$$T_{3,1;1,2} = 699.36$$

Accuracy of the Diffusion Approximation as a Function of Server 2 Utilization

Keeping the values of all the parameters constant, we changed the value of the service rates in center 2, such that the utilization of server 2 varies between 0.5 and .9. The results are compared with GPSS simulation composed of 100 blocks. Runs were performed for 10,000 transactions and consumed on the average 4.5 CPU minutes on the Sigma-9 computer system. We note that the confidence interval is widened as we increase the utilization. The reason is that for high utilization the number of samples (number of independent cycles) decreases. We can improve this by letting the simulation run for a larger number of transactions, however, the cost of computing is prohibitive, and we are satisfied to compare the results with the means for these cases.

Table 6.1 and figs. 6.2 and 6.3 summarize the results of this example. We find that the accuracy of the approximation improves with the increase of the utilization. We conclude by noting that the change of any parameter that tends to produce heavy traffic in one or more centers in the network will improve the accuracy of the diffusion approximations. Also, the accuracy of all the estimates is quite satisfactory for design purposes.

	$T_{1,1;3,2}$	Simulation	90% Conf.Int. for Mean	Diffusion	Relative Error %
$\rho_2 = .9$	$T_{1,1;3,2}$	721.12	459.29-883.90	699.36	3.02
	$T_{3,1;1,2}$	711.45	455.90-872.14	699.05	1.70
$\rho_2 = .75$	$T_{1,1;3,2}$	410.30	372.43-438.92	461.98	-12.60
	$T_{3,1;1,2}$	412.99	375.89-441.12	461.98	-11.86
$\rho_2 = .5$	$T_{1,1;3,2}$	341.95	299.11-375.08	390.09	-14.08
	$T_{3,1;1,2}$	358.08	308.32-395.65	390.09	- 8.94

Table 6.1 Accuracy vs. Utilization of Server 2

Fig. 6.2 Delay $T_{1,1;3,2}$ vs. Utilization of Server 2

Fig. 6.3 Delay $T_{3,1;1,2}$ vs. Utilization of Server 2

6.4.3 Example 2

Let us consider a second example in which one of the centers has an alternating priority service discipline. The network is shown in fig. 6.4 and consists of 3 service centers and each center has two classes of customers marked 1 and 2. Servers 1 and 3 have Preemptive-Resume service discipline, while server 2 has alternating priority discipline. This network is a model for two processors connected through a channel as will be shown in Chapter 7. External arrivals to the network are Poisson with mean rate λ , and the ratio R is assumed to be equal to 1. The states of the network are given by the pair (i, l) where the first number identifies the center and the second identifies the class. Class 1 in the Preemptive-Resume centers has the higher priority. Service time distributions are all negative exponential with the following means:

$$\begin{aligned} \mu_{1,1}^{-1} &= 4 & \mu_{1,2}^{-1} &= 32 \\ \mu_{2,1}^{-1} &= 36 & \mu_{2,2}^{-1} &= 36 \\ \mu_{3,1}^{-1} &= 2 & \mu_{3,2}^{-1} &= 20 \end{aligned}$$

The vector $\bar{q} \triangleq \{q_{i,l}\}$ is given by

$$\bar{q} = \{0.5, 0, 0, 0, .5, 0\}$$

Adding two nodes for the source and the sink, the transition matrix $\{r_{i,l;j,m}\}$ will be an (8 X 8) matrix and is given by

Fig. 6.4 Open Network of Example 2

$i, l \backslash j, m$	o	$1,1$	$1,2$	$2,1$	$2,2$	$3,1$	$3,2$	s
o	o	.5	o	o	o	.5	o	o
$1,1$	o	o	1	o	o	o	o	o
$1,2$	o	o	o	.5	o	o	o	.5
$2,1$	o	o	o	o	o	1	o	o
$2,2$	o	1	o	o	o	o	o	o
$3,1$	o	o	o	o	o	o	1	o
$3,2$	o	o	o	o	.5	o	o	.5
s	o	o	o	o	o	o	o	o

where states o and s represent the source and sink centers respectively.

Given $\Lambda^{-1} = 40$, we apply the algorithm as follows:

Step 1

We solve the set of simultaneous equations given by Eq. 6.27 to get:

$$e_{1,1} = e_{1,2} = e_{3,1} = e_{3,2} = 1$$

$$e_{2,1} = e_{2,2} = .5$$

Then, using Eq. 6.26 we find the mean arrival rates as

$$\lambda_{1,1} = \lambda_{1,2} = \lambda_{3,1} = \lambda_{3,2} = \Lambda = 1/40$$

$$\lambda_{2,1} = \lambda_{2,2} = .5\Lambda = 1/80$$

We notice that in this example we could obtain the above mean arrival rates by following the traffic in the network. However, in more complicated networks we have to use Eqs. 6.27 and 6.26.

The utilizations of the servers are calculated as:

$$\rho_1 = 4\lambda + 32\lambda = .9$$

$$\rho_2 = 36\lambda/2 + 36\lambda/2 = .9$$

$$\rho_3 = 2\lambda + 20\lambda = .55$$

Step 2

The departure processes for centers 1 and 3 are calculated from the service parameters for class 1, and from Eq. 3.17 for class 2.

Therefore, for center 1, $\beta_1 = .125$

$$c_d^{1,1} = 1$$

$$c_d^{1,2} = 0.827$$

and for center 3, we find

$$\beta_3 = 0.1$$

$$c_d^{3,1} = 1$$

$$c_d^{3,2} = .851$$

For center 2, we use Eqs. 5.9 and 5.12 and find

$$\beta_2 = 1$$

$$c_d^{2,1} = c_d^{2,2} = .5$$

Step 3

Using Eq. 6.38 we find

$$c_a^{1,1} = 1 + (.5 - 1) \times .5 = .75$$

$$c_a^{1,2} = 1 + (1 - 1) \times 1 = 1$$

$$c_a^{2,1} = 1 + (.827 - 1)(.5)^2/.5 = .914$$

$$c_a^{2,2} = 1 + (.851 - 1)(.5)^2/.5 = .926$$

$$c_a^{3,1} = 1 + (.5 - 1) \times .5 = .75$$

$$c_a^{3,2} = 1 + (1 - 1) \times 1 = 1$$

Step 4

The values of $\hat{\rho}_{i,l}$ are calculated as

i, l	1,1	1,2	2,1	2,2	3,1	3,2
$\hat{\rho}_{i,l}$	.187	.891	.794	.795	.160	.526

Step 5

The mean number of customers of class l in center i ($\bar{Q}_{i,l}$) is calculated as

(i, l)	1,1	1,2	2,1	2,2	3,1	3,2
$\bar{Q}_{i,l}$	.123	8.15	3.85	3.87	.06	1.11

Step 6

The mean time spent between (1,1) and (3,2) is calculated using Eq. 6.40 as follows:

$$\begin{aligned} T_{1,1;3,2} &= 40 \left\{ .123 + 8.15 + \frac{3.85}{.5} + .06 + 1.11 \right\} \\ &= 685.72 \end{aligned}$$

Similarly,

$$\begin{aligned} T_{3,1;1,2} &= 40 \left\{ .06 + 1.11 + \frac{3.87}{.5} + .123 + 8.15 \right\} \\ &= 687.32 \end{aligned}$$

Comparison with Simulation and Poisson Arrival Assumption

To check the results of the previous algorithm simulation runs, for 10,000 customers, were performed to obtain the means and the confidence intervals for the numbers in the system and the time delays. The simulation program was composed of approximately 140 GPSS blocks, and the CPU time was 4.4 minutes on the Carleton University Sigma-9 computer. Also for comparison, analytical expressions (that assume Poisson arrival) were used to calculate the means. The results are summarized in table 6.2. Though both the diffusion approximation and the analytical approximation gave satisfactory results, we notice that the diffusion approximation was superior in estimating $\bar{Q}_{2,1}$ and $\bar{Q}_{2,2}$ because it takes into consideration that the arrivals to these queues are not Poisson. As a result of this, the estimates of the times $T_{1,1;3,2}$ and $T_{3,1;1,2}$ were better for the diffusion approximation.

	Simulation	90% Conf.Int. for Mean	Diffusion	Analytical Approximation (Poisson arrival)
$\bar{Q}_{1,1}$	.112		.123	.111
$\bar{Q}_{1,2}$	8.26	6.18 9.57	8.15	8.11
$\bar{Q}_{2,1}$	3.51	2.5 4.17	3.85	4.5
$\bar{Q}_{2,2}$	3.54	2.65 4.1	3.87	4.5
$\bar{Q}_{3,1}$	.035		.06	.053
$\bar{Q}_{3,2}$	1.36	1.18 1.47	1.11	1.12
$T_{1,1;3,2}$	676.78	538.77 763.92	685.72	735.66
$T_{3,1;1,2}$	675.09	552.12 749.02	687.32	735.66

Table 6.2 Results of Example 2 for Open Networks

Accuracy of the Diffusion Approximation as a Function of Throughput

To investigate the effect of changing the throughput of the network on the accuracy of the approximation, we changed Λ (the total mean rate of arrival to the network) and obtained the means and the confidence intervals for the time spent in the network of example 2, using GPSS simulation. Also, we calculated the means using the diffusion approximation algorithm. The results are summarized in table 6.3, and figs. 6.5 and 6.6. These results show that the accuracy of the approximation improves with the increase of the throughput due to the heavy traffic assumption inherent in the diffusion approximation. However, the error in the low throughput case was around 20% which is still satisfactory for the initial phases of network design.

6.5 Diffusion Analysis of Closed Networks with Two Classes of Customers

6.5.1 Review of Literature on Closed Networks

As we mentioned earlier in this chapter, Gordon and Newell [GN67] developed a solution for the general closed networks with exponential service times and a single class. The solution showed that the joint equilibrium distribution has a product form. However, it has a normalizing constant that is a function of all possible states, and the calculation of this normalizing constant becomes tedious for systems with a large number of states (The number of the states is a function of the number of customers (n), and the number of centers (N) in the network). Buzen [Bu73] proposed an iterative algorithm for the calculation of the normalizing constant, which reduces the computational cost.

Throughput		Simulation	90% Conf. Int. for Mean	Diffusion	Relative Error %
$\Lambda = \frac{1}{40}$	$T_{1,1;3,2}$	676.78	538.77-763.92	685.72	- 1.32
	$T_{3,1;1,2}$	675.09	552.12-749.02	687.32	- 1.81
$\Lambda = \frac{1}{45}$	$T_{1,1;3,2}$	373.70	333.80-405.77	341.37	8.65
	$T_{3,1;1,2}$	391.21	344.58-429.60	342.27	12.51
$\Lambda = \frac{1}{50}$	$T_{1,1;3,2}$	296.00	274.52-313.52	239.62	19.05
	$T_{3,1;1,2}$	304.27	281.08-323.08	240.2	21.06

Table 6.3 Accuracy vs. Throughput for Example 2

Fig. 6.5 Delay $T_{1,1,3,2}$ vs. Throughput

$T_{3,1,1,2}$ Fig. 6.6 Delay $T_{3,1,1,2}$ vs. Throughput

Gordon and Newell [GN67] presented asymptotic analysis for balanced (approximately equal relative utilization for all servers) and unbalanced systems. Moore [Mo71] introduced the concept of a "bottleneck" in an unbalanced network (a bottleneck is the server with the largest relative utilization) and used Gordon and Newell's model to predict "bottlenecks" in computer systems. Muntz and Wong [MW74] extended the "bottleneck" analysis to networks with general service times at each node while maintaining that all customers are of one class.

Diffusion approximations were used in the literature to solve some cases of the closed networks. Gaver and Shedler [GS73] used the diffusion equation with two reflecting boundaries to describe the CPU utilization in a cyclic queue with general service times and FCFS queueing discipline in both servers. They showed that the asymptotic solution ($n \rightarrow \infty$) is a good approximation for the finite, yet large, values of n . The approach assumes an unbalanced system. A related work is the diffusion approximation of the repairman problem with two types of failure [GL77].

Kobayashi [Ko74] derives a diffusion approximation solution for the general network with N centers, general service time for each server, and FCFS service discipline. However, there is a normalizing constant in the joint distribution, and it is tedious to calculate for a network with a large number of states. The main problem in solving the general closed networks is to calculate the utilization of the different servers in the networks, and two approximate techniques can be used for this purpose. The first is based on the bottleneck analysis in unbalanced systems. The utilization of the bottleneck center will approach 1 as $n \rightarrow \infty$, therefore, we can estimate

the utilization of the rest of the centers. The second approach is to approximate the values for the utilization by assuming that the network has exponential servers of the "Gordon and Newell" type. The latter approach was used by Reiser and Kobayashi [RK74] in conjunction with the diffusion approximation for closed networks.

6.5.2 Closed Networks with Two Classes

In this section, we propose a solution for a general closed network with two classes of customers. We assume the network is unbalanced and one or more bottleneck(s) are formed, then we show that we can solve the rest of the network (unsaturated portion) as an open network with external arrivals.

Consider a network with finite number N of service centers, and there are two classes of customers circulating in the network. There are n_1 customers of class 1 and n_2 customers of class 2. A customer keeps his class as long as he stays in one center but may change classes after receiving service and before joining another center or returning to the same center. The transition probability $r_{i,l;j,m}$ is the probability that a customer of class l at center i will join class m at center j , upon receiving service at center i ($1 \leq i, j \leq N$ and $l, m = 1, 2$).

We assume also that we decomposed the network to irreducible Markov subchains. Let $\lambda_{i,l}$ denote the mean arrival rate of class l customers at center i , then

$$\lambda_{i,l} = \sum_{j=1}^N \sum_{m=1}^2 \lambda_{j,m} r_{j,m;i,l} \quad (6.41)$$

This set of linear equations has no unique solutions, but we can determine the relative mean arrival rates. In particular, we calculate the value of

$$e_{i,l} = \frac{\lambda_{i,l}}{\lambda_{1,1}} \quad 1 \leq i \leq N \text{ and } l = 1, 2 \quad (6.42)$$

where $e_{i,l}$ is the average number of visits to class l in center i for every visit to class 1 in center 1.

The relative utilization of center i is given by

$$\rho_i = \sum_{l=1}^2 \frac{\lambda_{i,l}}{\mu_{i,l}} = \lambda_{1,1} \sum_{l=1}^2 \frac{e_{i,l}}{\mu_{i,l}} \quad (6.43)$$

Since we cannot calculate a value for $\lambda_{1,1}$, therefore, we cannot estimate the values of the relative utilization of the different centers. However, if we assume the network is unbalanced, i.e. it has different utilization for the different centers, and $n_{1,2}$ ($n_{1,2} = n_1 + n_2$) is large enough, then the utilization of one or more of the centers will approach 1, and these centers will become bottlenecks. We find these centers by calculating the ratio $\rho_i/\lambda_{1,1}$, $1 \leq i \leq N$, and finding i that makes this ratio the largest. Without loss of generality, we assume one saturated center, which is the N^{th} center. As $n_{1,2} \rightarrow \infty$ the probability of having class 1 or class 2 customer in the saturated center will be α and $1 - \alpha$, respectively, where

$$\alpha = \frac{n_1}{n_{1,2}} \text{ and } n_{1,2} = n_1 + n_2 \quad (6.44)$$

Now, we open the network at the bottleneck and assume the departure process of this center is the external arrival to the rest of the centers ($1 \leq i \leq N - 1$).

We can then proceed to solve the network as an open network, with additional reflecting boundaries due to the finite number of customers that exist in the network from each class.

The result will be that we add the following condition

$$\sum_{n=0}^{n_l} \hat{p}_{i,l}(n) = \int_0^{n_l+1} \hat{p}_{i,l}(x) dx = 1 \quad l = 1, 2 \quad (6.45)$$

to the solution given by Eq. 3.34 to obtain

$$\hat{p}_{i,l}(n) = \frac{(1 - \hat{\rho}_{i,l}) \hat{\rho}_{i,l}^n}{1 - \hat{\rho}_{i,l}^{n_l+1}} \quad 0 \leq n \leq n_l \quad (6.46)$$

Notice that the limit of the integration in Eq. 6.45 is n_l+1 rather than n_l as a result of the discretization used which defines $\hat{p}_{i,l}(n)$ as

$$\hat{p}_{i,l}(n) = \int_n^{n+1} \hat{p}_{i,l}(x) dx$$

For the high priority class in a Preemptive-Resume, we know that

$\hat{p}_{i,1}(0) = 1 - \rho_{i,1}$, where $\rho_{i,l}$ is the relative utilization of server i by class l .

Therefore,

$$\hat{p}_{i,l}(n) = \begin{cases} 1 - \rho_{i,l} & n = 0 \\ \frac{\rho_{i,l} (1 - \hat{\rho}_{i,l}) \hat{\rho}_{i,l}^{n-1}}{1 - \hat{\rho}_{i,l}} & 1 \leq n \leq n_l \end{cases} \quad (6.47)$$

6.5.3 The Algorithm

The following algorithm is for two class closed networks:

1. Solve Eqs. 6.41, 6.42 and 6.44 to obtain $e_{i,l}$ and use those values to calculate ρ_i using Eq. 6.43, ρ_i will be a function of $\lambda_{1,1}$.
2. From the values of ρ_i , identify the bottleneck(s). Assume it is the N^{th} center and $\rho_N = 1$, and obtain the relative utilization of the rest of the centers ρ_i , $1 \leq i \leq N - 1$.
3. Using results in Chapters 2 to 6, obtain the departure process of center N which is considered as the external arrival to the network.
4. Solve for $c_d^{i,l}$ ($1 \leq i \leq N - 1$, $l = 1, 2$) using the results of the previous Chapters 2 to 5.
5. Obtain $c_a^{i,l}$ using Eq. 6.38.
6. For center i , calculate $\hat{\rho}_{i,l}$ given as

$$\hat{\rho}_{i,l} = \exp \left\{ \frac{-2(1 - \rho_i^*)}{\rho_i^* c_a^{i,l} + c_d^{i,l}} \times D_{i,l} \right\} \quad l = 1, 2 \quad (6.48)$$

where $D_{i,l}$ is calculated as shown in the algorithm for open networks.

7. Calculate the marginal probability distribution of finding n customers of class l in center i using Eqs. 6.46 or 6.47.

8. Calculate the mean number of customers of class l in center i

$$\begin{aligned}\bar{Q}_{i,l} &= \sum_{n=0}^{n_l} n \hat{p}_{i,l}(n) \\ &= \frac{\hat{\rho}_{i,l}}{1 - \hat{\rho}_{i,l}} - \frac{(n_l + 1) \hat{\rho}_{i,l}^{n_l+1}}{1 - \hat{\rho}_{i,l}^{n_l+1}} \quad \begin{matrix} 1 \leq i \leq N-1 \\ l = 1, 2 \end{matrix} \end{aligned} \quad (6.49)$$

Note that as $n_l \rightarrow \infty$, the second term vanishes.

For class 1 customers in a Preemptive-Resume queueing center the mean $\bar{Q}_{i,1}$ is obtained using Eq. 6.47 as

$$\bar{Q}_{i,1} = \frac{\rho_{i,1}}{1 - \rho_{i,1}} - \frac{1 - \hat{\rho}_{i,1}^{n_1+1}}{1 - \hat{\rho}_{i,1}^{n_1}} - \frac{\rho_{i,1} (n_1 + 1) \hat{\rho}_{i,1}^{n_1}}{1 - \hat{\rho}_{i,1}^{n_1}} \quad (6.50)$$

6.6 Numerical Results for Closed Networks

6.6.1 Introduction

In this section, we present two examples of unbalanced closed networks, and compare the results obtained from the previous algorithm with simulation. For the simulation runs, confidence intervals for the mean number of customers are obtained using the independent cycles' technique, where a cycle starts when all the customers in the network of the two classes are in the saturated center, and the rest of the servers are idle.

6.6.2 Example 3

Consider the simple cyclic network of fig. 6.7 which consists of two centers, and each center has two classes of customers (1 and 2). Server 1 gives priority to class 1 customers, while server 2 deals with all customers on FCFS basis. There are 10 customers of the first class, and 8 customers of the second class. This network represents a computer system with one CPU and one peripheral device (e.g. a disk), which handles time sharing and batch jobs. The high priority customers are the active members of the time sharing jobs, and we assume there is the same number of active members at any time, though they definitely belong to different time sharing customers. The number of the low priority customers represents the degree of multiprogramming in the system. The two classes of jobs are treated equally when they request I/O tasks from the peripheral device.

The service time distributions for all servers are exponential with the following rates of service for the two classes:

$$\mu_{1,1} = 1/10 \quad \mu_{1,2} = 1/30$$

$$\mu_{2,1} = \mu_{2,2} = 1/30$$

The transition matrix is a (4X4) matrix, and is given by

$i,l \backslash j,m$	1,1	1,2	2,1	2,2
1,1	0	0	1	0
1,2	0	0	0	1
2,1	1	0	0	0
2,2	0	1	0	0

Fig. 6.7 Cyclic Network of Example 3

We now apply the algorithm:

Step 1

From Eqs. 6.41 and 6.44

$$\lambda_{2,1} = \lambda_{1,1}$$

$$\lambda_{1,2} = \lambda_{2,2} = \frac{n_2}{n_1} \lambda_{1,1} = \frac{8}{10} \lambda_{1,1}$$

Therefore, from Eq. 6.43

$$\begin{aligned} \rho_1 &= \lambda_{1,1} \times 10 + \lambda_{1,2} \times 30 \\ &= 34 \lambda_{1,1} \end{aligned}$$

$$\text{and} \quad \rho_2 = (\lambda_{1,1} + \lambda_{2,1}) \times 30 = 54 \lambda_{1,1}$$

Step 2

Center 2 will become a bottleneck, and therefore

$$\rho_2 = 54 \lambda_{1,1} = 1 \quad \text{or} \quad \lambda_{1,1} = \frac{1}{54}$$

$$\text{and} \quad \rho_1 = \frac{34}{54} = 0.63$$

Step 3

Since service in center 2 is exponential, then

$$c_d^{2,1} = c_d^{2,2} = 1$$

Step 4

The unsaturated part of the network of this example is center 1 which has Preemptive-Resume service discipline, then

$$c_d^{1,1} = 1$$

and from Eq. 3.17

$$c_d^{1,2} = \frac{n_2^2 \frac{c_{1,2}}{\beta^2} + (n_2 c_{1,1} + 1)n_1}{(\frac{n_2}{\beta} + n_1)^2}$$

$$= .576$$

Step 5

From Eq. 6.38

$$c_a^{1,1} = c_a^{1,2} = 1$$

Step 6

The value of $\rho_{1,1}$ is given as

$$\rho_{1,1} = \frac{\lambda_{1,1}}{\mu_{1,1}} = \frac{10}{54} = .185$$

and from Eq. 6.48

$$\hat{\rho}_{1,1} = \exp \left\{ \frac{-2(1 - .185)}{1 + .185} \right\} = .253$$

and

$$\hat{\rho}_{1,2} = \exp \left\{ \frac{-2(1 - .63)}{.63 + .58} \right\} = 0.54$$

Step 7

From Eqs. 6.47 and 6.48, we obtain

n	0	1	2	3	4	5	6	7	8	9	10
$\hat{p}_{1,1}(n)$	.815	.138	.035	9×10^{-3}	2×10^{-3}	5×10^{-4}	1.3×10^{-4}	3×10^{-5}	8×10^{-6}	2×10^{-6}	5×10^{-7}
$\hat{p}_{1,2}(n)$	.46	.248	.134	.073	.040	.022	.012	.007	.004	-	-

Step 8

From Eqs. 6.50 and 6.49

$$\bar{Q}_{1,1} = .248$$

and

$$\bar{Q}_{1,2} = 1.139$$

Comparison with Simulation

GPSS simulation of the previous system was performed. The program has 50 blocks and was run for 20,000 simulation clock units, which took 0.3 CPU minutes. The results are summarized in table 6.4 and fig. 6.8, which show very satisfactory estimates of the means and the distributions.

		Queue 1	Queue 2
Simulation	mean	.244	1.189
	90% Conf. Inf.	[.202-.257]	[0.95-1.46]
Diffusion	mean	.248	1.139
	% error	-1.64	4.2

Table 6.4 Accuracy of Diffusion Approximation for Example 3

Accuracy of the Diffusion Approximation as a Function of Squared Coefficient of Variation $c_{1,2}$

We ran two more cases; the first case has Erlang-2 service distribution for the low priority customers in center 1, i.e. $c_{1,2} = .5$. The second has a 2-stage hyperexponential service distribution of the type introduced in Chapter 3 ($c_{1,2} = 1.5$).

Fig. 18 The Distribution of the Number of Customers in Center 1 of Example 3

Table 6.5 summarizes the results of these two cases and the previous one for the low priority customers. The results are in general satisfactory. However, the accuracy is better for the exponential system.

$c_{1,2}$	$\hat{\rho}_{1,2}$	Simulation mean	90% Conf.Int.	Diffusion mean	Relative Error %
.5	0.463	1.18	.9 - 1.41	0.853	27.7
1	0.54	1.189	.95- 1.46	1.139	4.2
1.5	.601	1.244	1.03- 1.42	1.413	-13.6

Table 6.5 Accuracy vs. $c_{1,2}$ for Example 3

6.6.3 Example 4

We consider here a central server model composed of 3 centers and 2 classes of customers as shown in fig. 6.9. Server 1 represents a CPU and gives priority to class 1 customers, while servers 2 and 3 represent I/O channels and handle all customers on an FCFS basis. In this example, we assume all service times to be exponentially distributed with the following means:

$$\mu_{1,1}^{-1} = 5 \qquad \mu_{1,2}^{-1} = 10$$

$$\mu_{2,1}^{-1} = \mu_{2,2}^{-1} = 20$$

$$\mu_{3,1}^{-1} = \mu_{3,2}^{-1} = 30$$

Fig. 6.9 Central Server Model of Example 4

Fig. 6.10 Reduced Model

The transition matrix is given by

$i, l \backslash j, m$	1,1	1,2	2,1	2,2	3,1	3,2
1,1	0	0	.5	0	.5	0
1,2	0	0	0	.5	0	.5
2,1	1	0	0	0	0	0
2,2	0	1	0	0	0	0
3,1	1	0	0	0	0	0
3,2	0	1	0	0	0	0

The number of class 1 customers in the network is 10 while there are 8 customers of the second class.

Solving Eqs. 6.41 and 6.44, we find

$$\lambda_{1,2} = 0.8\lambda_{1,1}$$

$$\lambda_{2,1} = \lambda_{3,1} = 0.5\lambda_{1,1}$$

$$\lambda_{2,2} = \lambda_{3,2} = .5\lambda_{1,2} = 0.4\lambda_{1,1}$$

then, from Eq. 6.43

$$\rho_1 = \lambda_{1,1} \times 5 + 0.8\lambda_{1,1} \times 10 = 13\lambda_{1,1}$$

$$\rho_2 = 0.9 \times 20\lambda_{1,1} = 18\lambda_{1,1}$$

$$\rho_3 = 0.9 \times 25\lambda_{1,1} = 22.5\lambda_{1,1}$$

Therefore, a bottleneck is formed in center 3 and its utilization will become 1, then

$$\lambda_{1,1} = 1/22.5,$$

$$\rho_1 = 0.578, \quad \text{and} \quad \rho_2 = .8$$

In the reduced model shown in fig. 6.10, the external arrivals $\lambda_{0,1}$ and $\lambda_{0,2}$ are given by

$$\lambda_{0,1} = \lambda_{3,1} = 1/45$$

$$\lambda_{0,2} = \lambda_{3,2} = 1/56.25$$

The results of this example are presented in table 6.6. In addition, simulation results are presented for comparison. Simulation was composed of 90 GPSS blocks, and was run for 200,000 simulation clock units, which consumed 3.5 CPU minutes on the Sigma-9 computer.

The accuracy of the diffusion approximation for this example is considered satisfactory. More errors are experienced with the FCFS center than the Preemptive-Resume center, and this is consistent with the results of the previous chapters.

i, l	$\hat{\rho}_{i,l}$	Simulation mean	90% Conf.Int.	Diffusion mean	Relative Error %
1,1	.28	.268	.260-.276	.308	-14.9
1,2	.461	.868	.828-.901	.855	1.5
2,1	0.641	2.119	1.677-2.443	1.703	19.6
2,2	0.592	1.803	1.444-2.069	1.370	23.9

Table 6.6 Accuracy of Diffusion Approximation for Example 4

6.7 Concluding Remarks

In this chapter, we presented two algorithms for the solution of open and closed networks with two classes of customers. We examined the accuracy of these algorithms by comparing their results with GPSS simulation. In all the cases studied, the accuracy is very satisfactory for network design purposes.

7. DESIGN OF FRONT-END AND HOST PROCESSOR INTERFACE CONFIGURATIONS

7.1 Introduction

In the last decade, there has been a rapid growth in teleprocessing services. The idea resulted from a substantial increase in computer usage and applications in large organizations, and a need to bring the computer power close to users and their environment. Teleprocessing systems allow several users direct access to the computer through remote consoles or terminals. Users can access and update data bases, enter and execute application programs, and send messages to other users in the system.

A teleprocessing system is a computer system that supports telecommunication functions as part of its operation. These functions are grouped together into what is referred to as the communication subsystem.

Examples of the main functions that are performed by the communication subsystem [M172] are:

1. Line Handling:

Line handling functions are needed to support transmission on a character-by-character basis without program intervention. Some of these functions concern assembling and disassembling characters, checking transmission errors, and recognizing special characters.

2. Network Control:

The network controller is responsible for handling all the interfaces with the telecommunication network. These include addressing and polling terminals, controlling devices and buffer management.

3. Message Editing:

The editor takes care of simple character and line deletion, in addition to interpreting message headers and adding control information (such as sequence number, date, time, etc.) to messages.

4. Message Queueing:

Some messages require queueing in main memory or mass storage device until the appropriate application program or console is available to receive them. These functions are executed under the supervision of the queueing controller, which will also signal the supervisory system to invoke the application program, or check the availability of the console.

5. Error Recovery:

In addition to the previous functions, the communication subsystem has a procedure for backup and graceful degradation of service in cases of hardware or software failures. Also it has recovery functions to protect itself from various network errors.

The implementation of the communication subsystem as part of the Host Processor leads to inefficient use of the processor's power. This inefficiency is due to the fact that a substantial amount of the Host Processor's CPU time is spent in executing the telecommunication functions while the processor itself is not designed to perform these functions efficiently. The development of relatively inexpensive and powerful mini-computers and multi-microprocessors suggested the use of small processors to perform the telecommunication functions, thus freeing the Host Processor to perform its main tasks efficiently. Such small processors are usually called Communication Processors (CP) or Front-End-Processors (FEP).

The designer of teleprocessing systems in which some of the telecommunication functions are performed by Communication Processors must address two important problems:

- (1) Selection of the hardware (CPU, storage and interface devices) for the configuration.
- (2) Distribution of the telecommunication functions between the Host Processor and the Communication Processor.

In this chapter, we address the problems of function partitioning between the FEP and the Host Processor, and the choice of a configuration which is "cost-effective" with respect to certain performance criterion. In sec. 7.2, a review of the different approaches taken to analyse the communication subsystem is presented. In sec. 7.3, we develop a queueing network model for an FEP connected to a Host Processor via a data transfer channel. The model allows the allocation of the different telecommunication functions to either one of the two processors. In sec. 7.4, we use the diffusion approximation algorithm for open queueing networks developed in Chapter 6 to analyse the communication subsystem model. First, we analyse the model of four different configurations to examine the effect of the throughput on the communication delay, and to predict sources of bottlenecks. We follow this analysis by considering a "cost-effectiveness" measure which can be used to select a suitable configuration for a given set of parameter values. A summary of the contents of this chapter is included in sec. 7.5.

7.2 A Review of the Approaches Used in Analysing Communication Processors

Several Host-FEP interface configurations have been developed or suggested [NR72, Yo74] and some of them are currently employed in practical

systems. These configurations differ according to the distribution of the telecommunication functions between the Host and the FEP. In one end of the spectrum of configurations we have systems in which the FEP emulates the hard-wired controller, and in the other end we have systems in which the FEP is responsible for nearly all the telecommunication functions discussed before. Between these two ends, several possibilities exist. Furthermore, the configurations are different in the hardware architecture [NR72].

Most of the designers of these systems concentrated on analysing the Communication Processor alone, to ensure it is not a potential bottleneck that may hinder future expansion of the system. Tirrell [Ti72] analysed the throughput of the SANDAC programmable Communication Processor. He assumed that the significant portion of the machine cycles will be used for processing the data, while only a small fraction of the available machine cycles is required for I/O transfers.

Spragins [Sp72] studied the behavior of a Communication Processor under extreme loading conditions. Similar to Tirrell, he assumed that the processing of the arriving characters (necessary to determine their disposition when they initially arrive) is the primary factor in limiting performance under extreme loads. GPSS simulation was used and simple design criteria for ensuring satisfactory performance without resorting to an extreme worst case design were developed. Communication Processors proved to have an extremely sharp saturation threshold which is very sensitive to minor variation in character processing times. Chang and Gorenstein [CG72] modelled a configuration in which a disk system is shared among several computers. Chang [Ch75] analysed the delay introduced by

the teleprocessing system. His main objective was to determine bottlenecks that may arise in the future operation of the system being designed. Chou et al [CF75] used simulation techniques also to analyse the delays in centralized computer communication systems with 3 hierarchical levels.

Other related studies include those of Pye and Arozullah [PA75] in which GPSS was used to simulate 6 configurations of single and dual processors (with and without FEP) and triple processor with FEP. Throughput and mean response time per transaction were used as criteria for comparison. Arokiaaray and Kamal [AK74] employed empirical formulas for modelling and analysing the central computer in a teleprocessing network configuration.

Chou, McGregor, and Kaczmarek [CM75, MC75] went beyond analysing systems for potential bottlenecks by developing a technique that uses simulation of a queueing model to represent the Communication Processor and its external environment (terminals, I/O devices, and Hosts). Their model allows changing the different parameters and studying their effects. However, the simulation model used in their study is quite extensive and the cost of running it can be prohibitive.

A comprehensive review of the research in the optimization problem of the general teleprocessing network area was provided by Green and Tang [GT73]. They surveyed the different analytical models used for the network resource allocation and the Host resource allocation.

7.3 Modelling the Communication Subsystem

In this section, we present a new model for the communication subsystem. The model takes into account two important considerations: (1) messages, originating from the different parts of the communication subsystem may

be fed back as they enter different servers before they finally exit from the system, and (2) partitioning of the telecommunication functions between the two processors can be regarded as a variable since it is possible to allocate most of these functions to either of the two processors.

Figure 7.1 shows the components of the teleprocessing system. The users, by means of their remote terminals, interact with their application programs or data bases via the communication subsystem. The remote terminals are connected to the communication subsystem via a network of communication lines. It should be noted that the separation of the three components in fig. 7.1 is a logical one since part of the communication subsystem will exist in the Host Processor and will be run under its supervisor (operating system).

We are interested here in modelling the communication subsystem to analyse its performance and study the different ways of function partitioning between the two processors. Therefore, we consider the components linked to the communication subsystem in fig. 7.1 as external sources and sinks for messages. The communication subsystem consists of several functions; some of these functions will reside in the Host Processor while the rest will reside in the FEP. Messages generated from the users as well as the messages generated from the application programs will constitute the external traffic of the communication subsystem.

Figure 7.2 shows an FEP connected to the Host Processor through a channel (examples of this FEP are the IBM3705 communication controller and the ARPA IMP). In this configuration, we have four servers handling

Fig. 7.1 Teleprocessing System

Fig. 7.2 Communication Subsystem

the message flow: (1) the FEP's CPU, (2) the Host's CPU, (3) the channel linking them, and (4) a mass storage device for messages that require queueing. Messages arrive with a mean rate of Λ /sec from users, and the application programs produce an average of R messages, for each message received. Messages, characterized by interarrival time, length, and priority, are received upon their arrival to the FEP by the line handler. The messages will be buffered until they are processed by the FEP's CPU. The processing time as well as the structure of the output messages that will be sent later to the Host Processor depend on the telecommunication functions that are allocated to and handled by the FEP. The FEP will then interrupt the Host to transfer a character, a message, or a block of messages according to the output format of the FEP. The Host Processor receives these characters (messages or blocks) and sends them for further processing in the message handler. Upon the completion of the telecommunication processing functions, the messages will leave the communication subsystem to the application programs (or the data bases), or they will be queued on the disk awaiting the application program to be free.

We now proceed to construct a queueing model for the four servers in the communication subsystem.

FEP

The FEP is modelled as a single server handling two priority tasks:

Priority 1: I/O transfer of data.

Priority 2: Background processing of blocks. The background may also be a multipriority if more than one type of message is received.

The mean rate of external arrival of messages is $\Lambda(1 + R)/\text{sec.}$, and the mean service rates of tasks type 1 and 2 are μ_{11} and μ_{12} , respectively. We assume the message length has exponential distribution with mean $\bar{l}_1$.

Figure 7.3 shows the FEP queueing model. We describe now the service time distribution of the high and low priority tasks. The high priority queue handles the service of messages upon arrival and before departure. External bits arriving at the FEP will be assembled into characters in line interface units, and when detected by the line scanner will get transferred to an incoming buffer under program control or cycle steal basis. The transfer requires $b_1 \text{ sec/ch}$ (second per character) of the CPU time. After a message receives processing in the CPU, it gets transferred to the channel adapter and from there to the Host Processor over the channel. The time taken for the transfer of characters from memory to channel adapter depends on the type of adapter (program controlled transfer or cycle steal transfer). A fully buffered channel transfers data between the channel buffer and memory on a time interleaved with the regular machine cycle, therefore, no CPU machine cycles are required. In general, we assume that the CPU time required for character transfer to the channel adapter is $b_2 \text{ sec/ch}$.

The output rate from the channel differs from the input rate to the FEP by a small amount, δ , per message which results from the transfer of additional command and control information. Now, we calculate the mean service time for I/O transfers of messages by combining the times

$$\bar{T}_{11} = \bar{l}_1 (b_1 + b_2) + \delta b_2 \quad (7.1)$$

Fig. 7.3 FEP Model

Fig. 7.4 Service Time Model

If we take the average values of b_1 , b_2 and δ , then the service time for I/O tasks has the same distribution as ℓ_1 (exponential) with mean $(b_1 + b_2)\bar{\ell}_1$ and a bias δb_2 . It is obvious that transfers in the opposite direction (from channel to memory then to line interfaces) have the same distribution.

The I/O transfers discussed so far consume a small portion of the CPU processing power. Examples given in [Ti71] show that even for a substantial data volume transmitted on high speed lines, the I/O consumed only 5.2% of the machine cycles. The major processing power of the FEP is used in background processing. This is divided into two groups:

1. Character processing

examples are error checking, code conversion, etc.

2. Block processing

examples are formatting, addressing, address identification, and chaining blocks into messages.

We now present a model for the service time of the background processing, which is suitable for studying function partitioning as will be shown later in this chapter.

We assume that we have a set of functions i , $1 \leq i \leq h$, to be performed by the FEP. Let π_i be the probability that a request of type i service will be made, and N_i the mean number of macroinstructions needed for this service. If X_F is the macroinstruction execution rate of the FEP, then the mean service time of type i service, performed by the FEP, is

$$\frac{1}{\mu_i} = \frac{N_i}{X_F} \text{ sec.} \quad (7.2)$$

The service time is modelled by the h -stages shown in fig. 7.4. This is the model we previously described in Chapter 3, and we have derived expressions for the mean and the second moment of the service time if all the stages are exponential (see Eqs. 3.53 and 3.54).

Notice that in general μ_i depends on the message length if the i^{th} stage involves character processing, and is independent of message length if it involves block processing.

Channel

The channel may be half or full duplex. The case of half duplex can be modelled by single server and two queues served on alternating priority basis [Ch74] as shown in fig. 7.5a. The full duplex channel is modelled by two separate queues and servers as shown in fig. 7.5b.

The rates μ_{21} and μ_{22} are the mean transmission rates for a message through the channel from FEP to Host and from Host to FEP, respectively. Since the messages are assumed to have exponentially distributed length, therefore, the service times of the channel are also exponential.

Host CPU

As mentioned before, our model is concerned only with the telecommunication functions in the Host and treats the application programs as external sources and sinks of the traffic. Messages will be flowing to the application programs from the CPU at a mean rate Λ/sec . The messages originated by the application programs can be regarded as input to the CPU at a mean rate RA/sec . In other words, the server whose model will be shown soon represents only the processing portion of the CPU allocated to the telecommunication functions. The model also assumes that the application programs have lower priority than all the telecommunication

(a) Half Duplex Channel

(b) Full Duplex Channel

Fig. 7.5 Models for Channel

Fig. 7.6 Software Organization for a Teleprocessing System

functions. Thus, the application programs do not interfere with the telecommunication functions. Figure 7.6 shows typical software organization for teleprocessing systems, e.g., IBM's TCAM (Telecommunication Access Method) which operates under OS360 in 360 or 370 IBM computer systems [Ma72]. The highest two priority levels are reserved for the basic and additional communication functions. In addition, there are two levels of priority for application programs, along with low level background (nonteleprocessing) programs.

We use the model proposed for the FEP to represent the telecommunication functions in the Host Processor, i.e. we represent them by a server with two priority queues and Preemptive-Resume service discipline. The model for the service time of the low priority queue is composed of stages as shown in fig. 7.4.

Figure 7.7 shows the model for the Host CPU, in which μ_{31} and μ_{32} are the mean service rates for a message in the high and low priority queues, respectively.

Host Disk

Some of the messages, after receiving service from the CPU are routed to a disk where they wait in a special queue for the appropriate application program. We assume the probabilities that a message going to an application program or to the channel will get queued on the disk are q_1 and q_2 , respectively. Figure 7.8 shows a model for 2 disk module drives and one controller. In this model, when a message seeks access to a disk module, it waits first in a queue for the desired disk module. When service is granted, the disk controller initializes the seek action (controller is free during this time). At the completion of the seek and when the controller

Fig. 7.7 Host CPU

Fig. 7.8 Model for Disk

is available, the transfer of the message begins after a rotational delay (latency time).

Incorporating the controller queueing time as an additional service time [Se66] which can be calculated from the machine repair mode, we find an equivalent model for the disk which is a GI/G/1 model. The service time for the disk has mean μ_4^{-1} and variance σ_4^2 which can be calculated as

$$\mu_4^{-1} = T_s + T_c + T_r \quad (7.3)$$

where T_s is the mean seek time

T_c is the mean controller queueing time

T_r is the mean controller service time

(Latency time + transfer time)

$$\text{and} \quad \sigma_4^2 = \sigma_s^2 + \sigma_c^2 + \sigma_r^2 \quad (7.4)$$

Seaman et al [Se66] derived the relation between the above means and variances and the parameters of the disk.

The Communication Subsystem Model

We combine now the queueing models for the various components to form a complete model for the communication subsystem. The complete model can be regarded as a queueing network with 4 servers as shown in fig. 7.9.

The servers representing the FEP and Host CPU have Preemptive-Resume discipline. The server representing the channel has alternating priority discipline and the server representing the disk has FCFS discipline. All service times have general distributions, and arrival to the network is of the renewal type.

Fig. 7.9 Communication Subsystem Model

7.4 A Communication Subsystem Example

7.4.1 Performance Analysis Study

In this section, we study the effect of changing the throughput on the delay encountered in the communication subsystem for an example of a teleprocessing system. In this example, messages from users have Poisson arrival with mean rate λ messages/sec, and the message length has an exponential distribution with mean length $\bar{l}_1 = 20$ character. The ratio of the outflow from the application programs to the inflow to the application programs is one, i.e. $R = 1$, and δ is assumed to be equal to zero (no control characters are added in the FEP). We have three telecommunications functions ($i = 1, 2, 3$), in addition to the basic communication function ($i = 0$). The basic communication function has the highest priority and represents I/O transfer of messages. The three low priority telecommunication functions, and the probability of requesting service of each are shown in table 7.1. These functions were described before in sec. 7.1. Also, we assume that all queues reside in the main memory (disk is not used).

i	π_i	function 'i'
1	0.7	Network controlling functions
2	0.6	Queueing controlling functions
3	0.5	Editing functions

Table 7.1 Telecommunication Functions

We study four FEP-Host configurations which range from the FEP responsible for line handling only and working under the control of the Host, to the FEP responsible for all the telecommunication functions listed above. We assume that the FEP has a 800 nanosecond cycle time, and the Host computer is twice as fast, i.e. it has 400 nanosecond cycle time. Using the above information and given the values of the variables N_i which represent the macroinstruction required for each function i , we estimate the mean service times for the different functions when performed on any of the two processors.

The values of $1/\mu_i$ ($i = 0,1,2,3$) are summarized in table 7.2 for the four configurations. It should be noted that the values in this table are estimates for the mean service times. In practice these values can be obtained from benchmark measurements.

	Function i	Configuration			
		I	II	III	IV
FEP	0	10	4	4	4
	1	-	20	20	20
	2	-	-	12	12
	3	-	-	-	20
Host	0	3	1	1	1
	1	10	-	-	-
	2	6	6	1	1
	3	10	10	10	-

all times are in m.sec.

Table 7.2 Mean Service Times for FEP and Host in 4 Different Configurations

In configuration I, the FEP is emulating a hardwired controller (e.g. emulating IBM 270X). The Host is experiencing many interrupts from the FEP, and the time required for I/O transfers per message is 10 m.sec. in FEP and 3 m.sec. in the Host. The remainder of the telecommunication functions are performed in the Host.

Configuration II employs the FEP as a Programmable Network controller (e.g. IBM3705 communication controller). Function 1 is moved to the FEP, where it has twice the previous mean time because of the difference in speed of the two processors. In this configuration, the Host is only interrupted when a message or a block is ready to be transferred from the FEP. The reduction in the number of interrupts is considered by changing the mean time of I/O transfers from 3 m.sec. to 1 m.sec.

In configuration III, the FEP is taking care of the queueing control functions. However, there is still a need for some buffering in the Host. This is indicated by the small service time assigned to function 2 in the Host.

Finally, in configuration IV, the FEP is responsible for all the telecommunication functions. In all these configurations, the FEP is connected to the Host via a channel. The average transmission time for a message is 20 m.sec.

Having described the model and its parameters, we proceed to use the diffusion approximation algorithm, described in sec. 6.3, for open networks. The algorithm estimates the total delay for messages from FEP to Host (Delay 1) and from Host to FEP (Delay 2).

We also examine the effect of changing the throughput λ on the delay in the communication subsystem of each configuration. Application of the

algorithm to the four configurations provides the results shown in fig. 7.10 which relate the throughput Λ to the total delay for messages from FEP to Host.

From fig. 7.10, we find:

- Configuration I has the smallest delay among the four configurations because in it the Host computer (which is the fast processor) is used for performing all the telecommunication functions. However, as we will see later, this does not represent the most "cost-effective" configuration when we consider a cost/performance measure for evaluating the four configurations.
- As we increase the throughput, saturation will occur in one or more of the centers, and the delay in this center becomes the dominant factor in the total delay. The saturated center in configurations I and II is the channel, while in configurations III and IV it is the FEP.
- For good performance (relatively small delays) of each of these configurations, throughput should be kept below the saturation values.

7.4.2 Cost-Effectiveness Study

We described in the previous section four configurations for a communication subsystem. They differ in the distribution of functions between the FEP and the Host. We have studied the effect of changing the throughput on the delay in the subsystem. In this section, we compare the four configurations at a fixed (operating) throughput based on a cost/performance measure that combines the cost and the performance

Fig. 7.10 Communication Delay as a Function of Throughput for 4 Different Configurations of Front-End and Host Processors Interface

of each configuration. The cost/performance will be such that it improves when either the cost or the average delay decrease. For the purpose of illustration, we consider the following "cost-effectiveness" formula:

$$\text{cost-effectiveness} = 1/(\text{cost})^i \times (\text{average delay})^j \quad (7.5)$$

where i and j can assume several possible values depending on the

weights assigned to the cost and the average delay

and average delay = (delay 1 + delay 2)/2 .

For simplicity, we consider $i = j = 1$.

Based on the nominal commercial price lists of similar computers, we take the cost of the Host computer to be 3 million dollars. The cost of the FEP in each of the four configurations is given in the following table.

Configuration	I	II	III	IV
FEP cost $\times 10^3$ dollars	100	200	300	400

The cost of the channel is relatively small and we assume it is included in the FEP cost.

The total system cost is assumed to be the cost of the FEP added to the utilized portion of the Host.

$$\text{Cost} = c_{\text{FEP}} + \rho_{\text{Host}} * c_{\text{Host}} \quad (7.6)$$

The results of the comparison for a throughput of 12.5 messages/sec are summarized in table 7.3.

Configuration	I	II	III	IV
cost 10^4 dollars	149.5	92	79.5	52
delay 1 m.sec.	58.8	87.7	111.6	272.9
delay 2 m.sec.	56.9	87.1	117.2	277.1
cost-effectiveness 10^{-5} /dollar.sec.	1.19	1.24	1.10	0.70

Table 7.3 Comparison between 4 FEP-Host Interface Configurations

Fig. 7.11 shows the "cost-effectiveness" measure as a function of the cost. It is clear in this example that configuration II is the most cost-effective configuration. As seen in this example, such a configuration is not necessarily the one with the least cost or the lowest delay for a given throughput.

7.5 Concluding Remarks

In this chapter, we have modelled the communication subsystem in a teleprocessing system by an open queueing network. The model is constructed in a way that permits the flexibility of moving the telecommunication functions from one processor (say FEP) to another (Host). Using the diffusion approximation technique developed in the previous chapter, we have studied the performance of the communication subsystem under varying

Fig. 7.11 Cost-Effectiveness for 4 Different Configurations of Communication Subsystems

throughput, for four FEP-Host configurations. In these configurations we employ different partitioning of the telecommunication functions between the two processors. Finally, a measure for the "cost-effectiveness" of the communication subsystem was introduced, and used in evaluating and comparing the four configurations for a given throughput.

8. CONCLUSIONS AND SUGGESTIONS FOR FUTURE WORK

8.1 Conclusions

In this thesis we have developed diffusion approximations for GI/G/1 queueing systems with two classes of customers and the following service discipline:

- FCFS with a different service time distribution for each class
- Preemptive-Resume
- Alternating priority

The accuracy of the approximation was examined by comparing the diffusion results with analytical expressions and GPSS simulation. The approximation gave the lowest error for a "symmetrical" Markov queueing system (Poisson arrival, exponential service times distributions, and equal rates of arrival and service for both classes) under heavy traffic. Empirical expressions have been proposed to improve the accuracy of other "non-symmetrical" systems for FCFS and alternating priority disciplines. With these improvements, the accuracy was found to be satisfactory in all cases.

In the area of open queueing networks, we have extended Reiser and Kobayashi's algorithm [RK74] to networks with two classes of customers and the service disciplines mentioned before. Examples of two different networks were presented to show the application of the new algorithm. Estimates of the mean number of customers of each class at every center, and estimates of the total delay in the network were examined and compared with GPSS simulation results and analytical solutions which

assume Poisson arrival. The diffusion approximation's accuracy was very satisfactory and superior to the Poisson assumption in estimating queues with non-Poisson arrival. The accuracy of the diffusion approximation was examined as a function of some of the network parameters, and found to be satisfactory in all cases considered.

For unbalanced closed networks with two classes of customers, we have proposed a method to obtain an equivalent open network for the unsaturated part, which can be solved by the previous algorithm. We examined the algorithm by applying it in two examples: (1) a cyclic network, and (2) a central server one. The accuracy, obtained by comparison with simulation, was satisfactory in all cases studied.

We concentrated all the previous discussions on queueing systems with two classes of customers to make the presentation clear; however, the generalization of all the previous results to multiclass cases (more than two) is straightforward.

We have developed a model for the communication subsystem in a teleprocessing system. The model describes the behavior of a Front-End Processor connected to a Host computer via a channel. It has the form of an open queueing network. The model for the service facility in both processors permits different ways of partitioning the functions between the two processors. The diffusion approximation algorithm for open networks was used to study an example, and to compare the performance of four configurations with different function partitioning. Finally, because comparison on the basis of performance only is incomplete, we have introduced the cost factor through a definition of "cost-effectiveness" measure, and we used it to compare between the four configurations.

In conclusion, the approximations reported in this thesis proved to combine the accuracy level and the simplicity of use required in designing and performance analysis of practical systems with many parameters (e.g. teleprocessing systems).

8.2 Recommendations for Future Work

This research proved the value of the diffusion approximation in providing solutions for many practical queueing systems. Several problems are still to be solved in the area. Among these problems are the following:

1. The solution of queueing systems and networks with finite waiting room. Newell's work [Ne75] is pioneering in this area, but very limited in application due to the complexity of the results. Emphasis must be put on obtaining simple solutions for these cases.

2. The problem of obtaining transient solutions for queueing systems is very important in time varying systems, and the diffusion approximations seem to be the only solution. Efficient algorithms must be found for obtaining the transient solutions for general networks with multiclass customers.

3. In the area of teleprocessing network design, the network "optimization" problem remains largely unsolved. In this thesis, we considered the related problem of function partitioning in the communication subsystem. It is earnestly hoped that the work of this thesis will pave the way towards finding a solution for the general problem of function distribution in a teleprocessing system in some "optimum" way.

REFERENCES

- AK74 Arokiaraj, R. and S. Kamal, A Tool for Selecting a Computer System for a Centralized Teleprocessing Network, pp.824-829, NTC 1974.
- Av65 Avi-Itzhak, B., W.L. Maxwell and L.W. Miller, Queueing with Alternating Priorities, Ops. Res., 13, pp.306-318, 1965.
- Ba75 Baskett, F., K.M. Chandy, R.R. Muntz and F.G. Palacios, Open, Closed, and Mixed Networks of Queues with Different Classes of Customers, J. ACM, 22, pp.248-260, April 1975,
- Bu56 Burke, P.J., The Output of a Queueing System, Ops. Res., 4, pp.699-704, 1956.
- Bu73 Buzen, J.P., Computational Algorithms for Closed Queueing Networks with Exponential Servers, Comm. of ACM, 16, pp.527-531, September 1973.
- Ch72 Chandy, K.M., The Analysis and Solutions for General Queueing Networks, Proc. 6th Annual Princeton Conf. on Information Sciences, Princeton Univ., 1972.
- Ch75a Chandy, K.M., U. Herzog and L. Woo, Parametric Analysis of Queueing Network Models, IBM J. Res. Develop., 19, pp.36-42, January 1975.
- Ch75b Chandy, K.M., U. Herzog and L. Woo, Approximate Analysis of General Queueing Networks, IBM J. Res. Develop., 19, pp.43-49, January 1975.
- Ch77 Chandy, K.M., J.H. Howard Jr. and D.F. Towsley, Product Form and Local Balance in Queueing Networks, J. ACM, 24, pp.250-263, April 1977.
- CG72 Chang, J.H. and S. Gorenstein, A Disk File System Shared by Several Computers in a Teleprocessing Environment, Symposium on Computer-Communications Networks and Teletraffic, Polytechnic Institute of Brooklyn, New York, pp.177-185, April 1972.
- Ch74 Chang, J.H., Configuration and Analysis of a Communications Processor, Proceedings of ICCO, Stockholm, 1974.
- Ch75 Chang, J.H., Terminal Response Times in Data Communications Systems, IBM J. Res. Develop., pp.272-282, May 1975.

- CF75 Chou, W., W. Frank and R. Van Slyke, A Simulation of Centralized Computer Communications Systems, IEEE Trans. on Comm., pp.994-1001, September 1975.
- CM75 Chou, W. and P.V. McGregor, A Unified Simulation Model for Communication Processors, Proc. Sym. on Computer Networks, Gaithersburg, pp.40-46, 1975.
- CM65 Cox, D.R. and H.D. Miller, The Theory of Stochastic Processes, John Wiley, 1965.
- Co62 Cox, D.R., Renewal Theory, Methuen, London, 1962.
- CI74 Crane, M.A. and D.I. Iglehart, Simulating Stable Stochastic Systems I: General Multiserver Queues, J. ACM, 21, pp.103-113, January 1974.
- Fi77 Fischer, M.J., Analysis and Design of Loop Service Systems via a Diffusion Approximation, Ops. Res., 21, pp.269-278, March-April 1977.
- Ga68 Gaver, D.P., Diffusion Approximations and Models for Certain Congestion Problems, J. Appl. Prob., 5, pp.607-623, 1968.
- GL77 Gaver, D.P. and J.P. Lehoczky, A Diffusion Approximation Solution for a Repairman Problem with Two Types of Failure, Management Science, 24, pp.71-81, September 1977.
- GS73 Gaver, D.P. and G.S. Shedler, Processor Utilization in Multi-programming Systems via Diffusion Approximation, Ops. Res., 21, pp.569-576, March-April 1973.
- Ge75 Gelenbe, E., On Approximate Computer System Models, J. ACM, 22, pp.261-269, April 1975.
- GP76 Gelenbe, E. and G. Pujolle, The Behaviour of a Single Queue in a General Queueing Network, Acta Informatica, 7, pp.123-136, 1976.
- GS72 Gittman, I.I. and A.V. Skorohod, Stochastic Differential Equations, Springer Verlag, 1972.
- GN67 Gordon, W.J. and G.F. Newell, Closed Queueing Systems with Exponential Servers, Ops. Res., 15, pp.254-265, 1967.
- GT73 Green, P.E. and D.T. Tang, Some Recent Developments in Teleprocessing System Optimization, IEEE INTERCON, Vol. 1, pp.1-7, 1973.
- GH74 Gross, D. and C.M. Harris, Fundamentals of Queueing Theory, John Wiley, 1974.

- HF77 Halachimi, B. and W.R. Franta, A Diffusion Approximate Solution to the G/G/k Queueing System, Comput. & Ops. Res., 4, pp.37-46, 1977.
- He75 Heyman, D.P., A Diffusion Model Approximation for the GI/G/1 Queue in Heavy Traffic, The Bell Sys. Tech. Journal, 54, pp.1637-1646, November 1975.
- JJ57 Jackson, J.R., Networks of Waiting Lines, Ops. Res., 5, pp.518-521, 1957.
- JJ63 Jackson, J.R., Jobshop-like Queueing Systems, Manag. Sc., 10, pp.131-142, October 1963.
- JR54 Jackson, R.R.P., Queueing Systems with Phase Type Service, Opl. Res. Q., 5, pp.109-120, 1954.
- Ja68 Jaiswal, N.K., Priority Queues, Academic Press, 1968.
- Ka68 Kabak, I.W., Stopping Rules for Queueing Simulations, Ops. Res., 16, pp.431-437, 1968.
- Ki64 Kimura, M., Diffusion Models in Population Genetics, J. Appl. Prob., 1, pp.177-232, 1964.
- KJ75 Kleijnen, J.P., Statistical Techniques in Simulation, Part II, Marcel Dekker, N.Y., 1975.
- Kl75 Kleinrock, L., Queueing Systems, Volume I: Theory, John Wiley, 1975.
- Kl76 Kleinrock, L., Queueing Systems, Volume II: Computer Applications, John Wiley, 1976.
- Ko74 Kobayashi, H., Application of the Diffusion Approximation to Queueing Networks, I: Equilibrium Queue Distributions. J. ACM 21, pp.316-328, April 1974.
- Li61 Little, J.D.C., A Proof for the Queueing Formula, $L = \lambda w$, Ops. Res., 9, pp.383-387, 1961.
- Ma72 Martin, J., Systems Analysis for Data Transmission, Prentice Hall, 1972.
- MC75 McGregor, P., W. Chou and R. Kaczmarek, Communications Processor Simulation: A Practical Approach, NTC, New Orleans, pp.24-26 to 24-33, 1975.
- Mi72 Mills, D.L., Communication Software, Proc. IEEE, 60, pp.1333-1341, November 1972.

- Mo71 Moore, C.G. III, Network Models for Large-scale Time-sharing Systems, Tech. Rep. 71-1, Dept. of Ind. Eng., U. of Michigan, Ann Arbor, Mich., April 1971.
- Mo65 Morse, P.M., Queues, Inventories and Maintenance, Pub. in Ops. Res., John Wiley, 5th Edition, 1965.
- Mu73 Muntz, R.R., Poisson Departure Processes and Queueing Networks, Proc. of 7th Annual Princeton Conf. on Information Sciences and Systems, 1973.
- MW74 Muntz, R.R. and J.W. Wong, Asymptotic Properties of Closed Queueing Network Models, Proc. of the 8th Annual Princeton Conf. on Information Sciences and Systems, March 1974.
- Ne68 Newell, G.F., Queues with Time-Dependent Arrival Rates, Part I, pp.436-451; Part II, pp.579-590; Part III, pp.591-606, J. App. Prob., 5, 1968.
- Ne71 Newell, G.F., Applications of Queueing Theory, Chapman and Hall, London, 1971.
- Ne73 Newell, G.F., Approximate Stochastic Behavior of n-Server Service Systems with Large n, Springer Verlag, 1973.
- Ne75 Newell, G.F., Approximate Behavior of Tandem Queues, Special Report, Institute of Transportation and Traffic Engineering, Univ. of California, Berkeley, Oct. 1975.
- NR72 Newport, C.B. and J. Ryzlak, Communication Processors, Proc. IEEE, 60, pp.1321-1332, November 1972.
- Pa65 Papoulis, A., Probability, Random Variables, and Stochastic Processes, McGraw-Hill, 1965.
- Pi58 Pipes, L.A., Applied Mathematics for Engineers and Physicists, McGraw-Hill, 1958.
- PA75 Pye, W.T. and M. Arozullah, Load Distribution in Enquiry-Response Systems with Multiple Processors, ICC, pp.33-22 to 33-26, 1975.
- RK74 Reiser, M. and H. Kobayashi, Accuracy of the Diffusion Approximation for Some Queueing Systems, IBM J. Res. Develop., pp.110-124, March 1974.
- Sa75 Sauer, C.H., Configuration of Computing Systems: An Approach Using Queueing Network Models, Ph.D. Dissertation, U. of Texas at Austin, May 1975.

- SC75 Sauer, C.H. and K.M. Chandy, Central Server Model for Multi-programmed Computer Systems with Different Classes of Jobs, IBM J. Res. Develop., 19, pp.301-313, May 1975.
- Sc67 Scherr, A.A., An Analysis of Time-shared Computer Systems, MIT Press, Cambridge, Mass., 1967.
- Sc74 Schriber, T.J., Simulation Using GPSS, John Wiley, 1974.
- Se66 Seaman, P.H., R.A. Lind and T.L. Wilson, On Teleprocessing System Design, Part IV, An Analysis of Auxiliary-Storage Activity, IBM Systems Journal, 5, pp.158-170, 1966.
- Sp72 Spragins, J.D., Simulation of a Communication Processor Under Extreme Loading Conditions, IEEE, Trans. on Comm., 20, pp.609-619, June 1972.
- Ta68 Takacs, L., Two Queues Attended by a Single Server, Ops. Res., 16, pp.639-650, 1968.
- TN77 Taube-Netto, M., Two Queues in Tandem Attended by a Single Server, Ops. Res., 25, pp.140-147, 1977.
- Ti72 Tirrell, J.C., Analysis of Throughput of a Communication Processor, Computer Design, Part I, pp.47-51, September 1972, and Part II, pp.89-93, October 1972.
- WR66 Wallace, V.L. and R.S. Rosenberg, Markovian Models and Numerical Analysis of Computer System Behavior, Proc. Spring Joint Comp. Conf., 1966.
- Yo74 Yourdon, E., Design of On-line Computer Systems, Prentice Hall, 1972.

APPENDIX A

Table A.1 gives the definitions of the parameters used in the block diagram of the GPSS simulation for a Preemptive-Resume queueing system with two classes of customers. Figure A.1 shows the block diagram which is self explanatory. The flowchart symbols follow Schriber's terminology. [Sc74].

Table A.1

Table of DefinitionsTransactions

Model Segment 1	High-priority customers
Model Segment 2	Low-priority customers
Model Segment 3	A transaction to sample the number of customers in the system of each class every 500 time units

Functions

1	Exponential
2	Uniform

Savevalues

3	The sum of queueing times of all class 2 customers between regeneration points
4	The number of customers of class 2 who arrived at the system between regeneration points
6	The sum of variable 2

Tables

7	The distribution of the number of customers of class 1 in the system sampled every 500 time units
10	The distribution of savevalue 3 sampled at regeneration points
11	The distribution of savevalue 4 sampled at regeneration points
12	The distribution of the number of customers of class 2 in the system sampled every 500 time units

Variables

2	Savevalue 3 * savevalue 4
---	---------------------------

SEGMENT 1

SEGMENT 2

Fig. A.1 GPSS Block Diagram for a Preemptive-Resume Queuing System with Two Classes of Customers

Fig. A.1 (Continued)

APPENDIX B

The following table and block diagram describe the GPSS simulation of an FCFS queueing system with two classes of customers.

Table B.1

Table of DefinitionsTransactions

Model Segment 1	Class 1 customers
Model Segment 2	Class 2 customers
Model Segment 3	A transaction to sample the number of customers in the system of each class every 1000 time units

Functions

1	Exponential
---	-------------

Savevalues

1	The sum of the queueing times of all class 1 customers between regeneration points
2	The number of customers of class 1 who arrived at the system between regeneration points
3	Similar to savevalue (1) for class 2
4	Similar to savevalue (2) for class 2
5	The sum of variable 1
6	The sum of variable 2

Tables

1	The distribution of savevalue 1 sampled at regeneration points
5	The distribution of savevalue 2 sampled at regeneration points
7	The distribution of the number of customers of class 1 in the system sampled every 1000 time units

Tables (cont'd)

10

The distribution of savevalue 3 sampled at
regeneration points

11

The distribution of savevalue 4 sampled at
regeneration points

12

The distribution of the number of customers
of class 2 in system sampled every 1000 time
units

Variables

1

Savevalue 1 * savevalue 2

2

Savevalue 3 * savevalue 4

Fig. B.1 GPSS Block Diagram for an FCFS Queuing System with Two Classes of Customers

Fig. B.1 (Continued)

APPENDIX C

The following table and block diagram describe the GPSS simulation of an alternating priority queueing system with two classes of customers.

Table C.1

Table of DefinitionsTransactions

Model Segment 1	Class 1 customers
Model Segment 2	Class 2 customers
Model Segment 3	A transaction to sample the number of customers in the system of each class every 500 time units
Model Segment 4	A dummy transaction to initialize the system after regeneration points

Functions

1	Exponential
---	-------------

Logical Switches

LOCK	A switch to transfer service between the 2 classes (Reset for class 1, set for class 2)
CNTR	A switch to collect information and transfer service to the class with the 1 st arrival after regeneration points

Savevalues

1	The sum of queueing times of all class 1 customers between regeneration points
2	The number of customers of class 1 who arrived at the system between regeneration points
3	Similar to savevalue 1, for class 2 customers
4	Similar to savevalue 2, for class 2 customers

Savevalues (cont'd)

- 5 The sum of variable 1 calculated at regeneration points
- 6 The sum of variable 2 calculated at regeneration points

Tables

- 1 The distribution of savevalue 1 sampled at regeneration points
- 5 The distribution of savevalue 2 sampled at regeneration points
- 7 The distribution of the number of customers of class 1 in the system sampled every 500 time units
- 10 The distribution of savevalue 3 sampled at regeneration points
- 11 The distribution of savevalue 4 sampled at regeneration points
- 12 The distribution of the number of customers of class 2 in the system sampled every 500 time units

Variables

- 1 Savevalue 1 * savevalue 2
- 2 Savevalue 3 * savevalue 4

SEGMENT 1

SEGMENT 2

Fig. C.1 GPSS Block Diagram for an Alternating Priority Queueing

SEGMENT 4

END

05

0

79

FIN