



UNIVERSITI PUTRA MALAYSIA

**ANALYSIS AND DECENTRALISED OPTIMAL FLOW CONTROL
OF HETEROGENEOUS COMPUTER COMMUNICATION
NETWORK MODELS**

KU RUHANA KU MAHAMUD

FPP 1993 5

**ANALYSIS AND DECENTRALISED OPTIMAL FLOW CONTROL
OF HETEROGENEOUS COMPUTER COMMUNICATION
NETWORK MODELS**

By

KU RUHANA KU MAHAMUD

**Dissertation Submitted in Fulfilment of the Requirements for
the Degree of Doctor of Philosophy in the Faculty of
Science and Environmental Studies,
Universiti Pertanian Malaysia**

December 1993



ACKNOWLEDGEMENTS

I owe deep gratitude to Assoc. Prof. Dr. Abu Talib Othman, Chairman of the supervisory committee, for suggesting the problem to me and for his constant guidance and continuous encouragement without which this work would not have materialised.

I am also thankful to co-supervisors, Dr. Ali Mamat and Assoc. Prof. Dr. Borhanuddin Mohd. Ali for their kind assistance and scholarly guidance.

I am indebted to the Government of Malaysia and Universiti Utara Malaysia for sponsoring this research.

I am also deeply grateful to my husband and my children for their patience, support and care.



TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS	iii
LIST OF TABLES	viii
LIST OF FIGURES	ix
LIST OF ABBREVIATIONS	xvii
ABSTRACT	xix
ABSTRAK	xxii
CHAPTER	
I. INTRODUCTION	1
Background of the Problem	5
Objectives of the Study	9
Significance of the Study	10
Organisation of the Thesis	12
 II. MAXIMUM ENTROPY OF GENERAL QUEUEING NETWORK MODELS	 14
Introduction	14
Review of Related Literature	14
Single Server General Queueing Networks with Multiclass Users	15
Multiserver General Queueing Networks with Multiclass Users	17
Large General Queueing Networks with Single Class Users, Single Server and/or Multiserver Centres	18
Flow Control in Communication Networks	19
Maximum Entropy in Queueing Theory	21



	The Principle of Maximum Entropy	23
	Norton's Theorem in Queueing Networks	26
	Summary	28
III.	NORTON-MAXIMUM ENTROPY METHOD FOR LARGE GENERAL QUEUEING NETWORKS	30
	Introduction	30
	The General Queueing Model	31
	Service Station Aggregation	35
	Norton-Maximum Entropy Algorithm	39
	Numerical Examples	39
	Summary	51
IV.	SINGLE SERVER GENERAL QUEUEING NETWORKS WITH MULTICLASS USERS	52
	Introduction	52
	Description of the Network	53
	Maximum Entropy Formulations	54
	One-Step Maximum Entropy Recursions	66
	Extension to Load Dependent Queueing Systems	72
	Numerical Examples	74
	Summary	84
V.	MULTISERVER GENERAL QUEUEING NETWORKS WITH MULTICLASS USERS	85
	Introduction	85
	Description of the Network	86
	Maximum Entropy for Multiserver General Queueing System with Multiclass Users	86

	One-Step Maximum Entropy Recursions	103
	Extension to Load Dependent Queueing Systems	110
	Numerical Examples	112
	Summary	120
VI.	DECENTRALISED OPTIMAL FLOW CONTROL OF SINGLE SERVER GENERAL QUEUEING NETWORKS WITH MULTICLASS USERS	122
	Introduction	122
	The Mathematical Model	124
	The Optimisation Criterion	129
	The Optimal Control	136
	Numerical Examples	148
	Summary	157
VII.	DECENTRALISED OPTIMAL FLOW CONTROL OF MULTISERVER GENERAL QUEUEING NETWORKS WITH MULTICLASS USERS	159
	Introduction	159
	The Mathematical Model	151
	The Optimisation Criterion	167
	The Optimal Flow Control of a GE/GE/ ∞ Finite Capacity Queue	171
	The Optimal Flow Control of a GE/GE/c Finite Capacity Queue	177
	Numerical Examples	194
	Summary	206

VIII. CONCLUSIONS AND FUTURE RESEARCH	207
General Summary	207
Suggestion for Future Research	210
 BIBLIOGRAPHY	 213
APPENDICES	220
VITA	230



LIST OF TABLES

Table		Page
1	Comparison between Simulation (S), Maximum Entropy (ME) and Norton-ME (N-ME) for $M = 8$, $\mu_i = 8$, $i = 2, 3, \dots, 8$, $C_i^2 = 2$, $i = 1, 2, \dots, 8$	42
2	Comparison between Simulation (S), Maximum Entropy (ME) and Norton-ME (N-ME) for $M = 6$, $\mu_2 = 8$, $\mu_3 = \mu_4 = \mu_5 = \mu_6 = 10$, $C_1^2 = 1$, $C_2^2 = 0.5$, $C_3^2 = C_4^2 = C_5^2 = C_6^2 = 5$	44
3	Comparison between Simulation (S), Maximum Entropy (ME) and Norton-ME (N-ME) for $M = 7$, $\mu_3 = 8$, $\mu_i = 10$, $i = 1, 2, 4, \dots, 7$, all $C_i^2 = C_s^2$, $i = 1, 2, \dots, 7$	46
4	Comparison between Simulation (S_1, S_2) and Norton-ME (N-ME) for $M = 4$, $\mu_1 = 7$, $\mu_2 = \mu_3 = 10$, $C_i^2 = C_s^2$, $i = 1, 2, 3, 4$	48
5	Comparison between Simulation (S_1, S_2) and Norton-ME (N-ME) for $M = 6$, $\mu_1 = 8$, $\mu_i = 10$, $i = 2, 3, \dots, 5$, $C_i^2 = C_s^2$, $i = 1, 2, \dots, 6$	50

LIST OF FIGURES

Figure		Page
1	The Structure of a Computer Communication Network	3
2	Typical Plot of Throughput versus Offered Load	7
3	Typical Plot of Time Delay versus Offered Load	7
4	Queueing Network	27
5	Equivalent Queueing Network	27
6	Subsystem 1 Shorted	28
7	A General Queueing Network with M Service Centres	32
8	A "Shorted" Subnetwork 1 (σ) Cyclic Queueing Network	35
9	A "Shorted" Subnetwork 2 (σ^*) Cyclic Queueing Network	36
10	A Flow Equivalent Load Dependent Two Stage Cyclic Queueing System	37
11	A Closed Queueing Network with Fixed Routing for $M = 8$, $\mu_1 = 8$, $i = 1, 2, \dots, 8$, $C_1^2 = 2$, $i = 1, 2, \dots, 8$	41
12	Time Delay versus Number of Users	42
13	A Closed Queueing Network for $M = 6$, $\mu_2 = 8$, $\mu_3 = \mu_4 = \mu_5 = \mu_6 = 10$, $C_1^2 = 1$, $C_2^2 = 2$, $C_3^2 = C_4^2 = C_5^2 = C_6^2 = 5$	43
14	Throughput versus Number of Users	43



15	A Closed Queueing Network for $M = 7$, $\mu_3 = 8$, $\mu_1 = 10$, $i = 1, 2, 4, \dots, 7$, $C_1^2 = C_s^2$, $i = 1, 2, \dots, 7$	45
16	Time Delay versus Number of Users	45
17	A Closed Queueing with Fixed Routing for $M = 4$, $c = 3$, $\mu_1 = 7$, $\mu_2 = \mu_3 = 10$, $C_1^2 = C_s^2$, $i = 1, 2, \dots, 4$	47
18	A Flow Equivalent Load Dependent Two Stage Multiserver Cyclic Queueing System	47
19	A Closed Queueing Network with Fixed Routing for $M = 6$, $c = 2$, $\mu_1 = 8$, $\mu_1 = 10$, $i = 2, 3, \dots, 5$, $C_1^2 = C_s^2$, $i = 1, 2, \dots, 6$	49
20	A Flow Equivalent Load Dependent Two Stage Multiserver Cyclic Queueing System	49
21	Time Delay versus Number of Users	51
22	Queueing Model for Multiclass Users and Single Server Centres	53
23	Transition Weight Diagram of a $G/G/1/N$ Queue with Single Class Users	58
24	Transition Weight Diagram of a $G/G/1/N$ Queue with Two Classes of Users	58
25	Transition 'Up' and 'Down' Rate of a $GE/GE/1/N$ with Two Classes of Users	69
26	Queueing Model for Two Classes of Users and Load Dependent Centres	73
27	Throughput of Class 1 and 2 users versus Total Offered Load ($N_1 = 5$, $N_2 = 10$)	75
28	Mean Queue Length of Class 1 and 2 Users versus Total Offered Load ($N_1 = 5$, $N_2 = 10$)	76
29	Overall Throughput versus Total Offered Load ($N_1 = 5$, $N_2 = 10$)	76

30	Overall Mean Queue Length versus Total Offered Load ($N_1 = 5, N_2 = 10$)	77
31	Throughput of Class 1, 2 and 3 Users versus Total Offered Load ($N_1 = 5, N_2 = N_3 = 5$)	77
32	Mean Queue Length of Class 1, 2 and 3 Users versus Total Offered Load ($N_1 = 5, N_2 = N_3 = 5$)	78
33	Overall Throughput versus Total Offered Load ($N_1 = 5, N_2 = N_3 = 5$)	78
34	Overall Mean Queue Length versus Total Offered Load ($N_1 = 5, N_2 = N_3 = 5$)	79
35	Virtual Circuit Link Model for Two Classes of Users ($C_0^2 = C_{1a}^2$ $= C_{1b}^2 = C_{2a}^2 = C_{2b}^2 = C_{2c}^2 = 0.5$)	79
36	Throughput of Class 1 and 2 Users versus Total Offered Load	80
37	Mean Queue length of Class 1 and 2 Users versus Total Offered Load	80
38	A Queueing Model for Virtual Circuit Link with two Classes of Users	81
39	Throughput of Class 1 and 2 Users versus Total Offered Load	82
40	Mean Queue Length of Class 1 and 2 Users versus Total Offered Load	82
41	Queueing Model for Multiclass Users with Single Server and Multiserver Queueing Centres	87
42	Transition Weight of a G/G/2 Queue with Two Classes of Users	92
43	Transition Weight of a G/G/3 Queue with Two Classes of Users	95
44	Transition Weight of a G/G/c Queue with Two Classes of Users	98
45	Transition Weight of a G/G/c/N Queue with Two Classes of Users	100

46	Transistion 'Up' and 'Down' Rate of a GE/GE/c/N Queue with Two Classes of Users	106
47	Queueing Model for Two Classes of Users and Load Dependent Centres	111
48	Throughput of Class 1 and 2 Users versus Total Offered Load (G/G/2, $C_{\mu}^2 = 0.5$, $C_{\lambda_1}^2 = C_{\lambda_2}^2 = 2.0$)	113
49	Mean Queue Length of Class 1 and 2 Users versus Total Offered Load (G/G/2, $C_{\mu}^2 = 0.5$, $C_{\lambda_1}^2 = C_{\lambda_2}^2 = 2.0$)	113
50	Throughput of Class 1 and 2 Users versus Total Offered Load (G/G/3, $C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 0.5$)	114
51	Mean Queue Length of Class 1 and 2 Users versus Total Offered Load (G/G/3, $C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 0.5$)	114
52	Mean Response Time of Class 1 and 2 Users versus Total Offered Load (G/G/4, $C_{\mu}^2 = 0.5$, $C_{\lambda_1}^2 = C_{\lambda_2}^2 = 2.0$)	115
53	Mean Queue Length of Class 1 and 2 Users versus Total Offered Load (G/G/4, $C_{\mu}^2 = 0.5$, $C_{\lambda_1}^2 = C_{\lambda_2}^2 = 2.0$)	115
54	Throughput of Class 1 and 2 Users versus Total Offered Load (G/G/2/N, $N_1 = 5$, $N_2 = 10$, $C_{\mu}^2 = 0.5$, $C_{\lambda_1}^2 = 2.0$, $C_{\lambda_2}^2 = 1.0$)	116
55	Mean Queue Length of Class 1 and 2 Users versus Total Offered Load (G/G/2/N, $N_1 = 5$, $N_2 = 10$, $C_{\mu}^2 = 0.5$, $C_{\lambda_1}^2 = 2.0$, $C_{\lambda_2}^2 = 1.0$)	116



56	Throughput of Class 1 and 2 Users versus Total Offered Load (G/G/3/N, $N_1 = 5$, $N_2 = 10$, $C_\mu^2 = 0.5$, $C_{\lambda_1}^2 = C_{\lambda_2}^2 = 2.0$)	117
57	Mean Queue Length of Class 1 and 2 Users versus Total Offered Load (G/G/3/N, $N_1 = 5$, $N_2 = 10$, $C_\mu^2 = 0.5$, $C_{\lambda_1}^2 = C_{\lambda_2}^2 = 2.0$)	117
58	Throughput of Class 1 and 2 Users versus Total Offered Load (G/G/4/N, $N_1 = N_2 = 5$, $C_\mu^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 1.0$)	118
59	Mean Queue Length of Class 1 and 2 Users versus Total Offered Load (G/G/4/N, $N_1 = N_2 = 5$, $C_\mu^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 1.0$)	118
60	Overall Throughput versus Total Offered Load (G/G/4/N, $N_1 = N_2 = 5$)	119
61	Overall Mean Queue length versus Total Load (G/G/4/N, $N_1 = N_2 = 5$)	119
62	Model for Bottleneck Decentralised Flow Control	125
63	Transition 'Up' and 'Down' Rate of GE(n)/GE(n)/1/N with Two Classes of Users	128
64	Throughput versus Control ($C_\mu^2 = C_{\lambda_1}^2 =$ $C_{\lambda_2}^2 = 0.5$, $c_1 = 0.3$, $c_2 = 0.6$, $\mu = 1$)	150
65	Throughput versus Control ($C_\mu^2 = C_{\lambda_1}^2 =$ $C_{\lambda_2}^2 = 1.0$, $c_1 = 0.3$, $c_2 = 0.6$, $\mu = 1$)	150

66	Throughput versus Control ($C_{\mu}^2 = C_{\lambda_1}^2 =$ $C_{\lambda_2}^2 = 2.0, c_1 = 0.3, c_2 = 0.6, \mu = 1$)	151
67	Throughput versus Control ($C_{\mu}^2 = C_{\lambda_1}^2 =$ $C_{\lambda_2}^2 = 5.0, c_1 = 0.3, c_2 = 0.6, \mu = 1$)	151
68	Throughput versus Control ($C_{\mu}^2 = C_{\lambda_1}^2 =$ $C_{\lambda_2}^2 = 5.0, c_1 = 0.3, c_2 = 0.4, \mu = 1$)	152
69	Throughput versus Control ($C_{\mu}^2 = C_{\lambda_1}^2 =$ $C_{\lambda_2}^2 = 5.0, c_1 = 0.3, c_2 = 0.2, \mu = 1$)	152
70	Time Delay versus Control ($C_{\mu}^2 = C_{\lambda_1}^2 =$ $C_{\lambda_2}^2 = 0.5, c_1 = 0.3, c_2 = 0.6, \mu = 1$)	153
71	Time Delay versus Control ($C_{\mu}^2 = C_{\lambda_1}^2 =$ $C_{\lambda_2}^2 = 1.0, c_1 = 0.3, c_2 = 0.6, \mu = 1$)	153
72	Time Delay versus Control ($C_{\mu}^2 = C_{\lambda_1}^2 =$ $C_{\lambda_2}^2 = 2.0, c_1 = 0.3, c_2 = 0.6, \mu = 1$)	154
73	Time Delay versus Control ($C_{\mu}^2 = C_{\lambda_1}^2 =$ $C_{\lambda_2}^2 = 5.0, c_1 = 0.3, c_2 = 0.6, \mu = 1$)	154

74	Time Delay versus Control ($C_{\mu}^2 = C_{\lambda_1}^2 =$ $C_{\lambda_2}^2 = 5.0, c_1 = 0.3, c_2 = 0.4, \mu = 1$)	155
75	Time Delay versus Control ($C_{\mu}^2 = C_{\lambda_1}^2 =$ $C_{\lambda_2}^2 = 5.0, c_1 = 0.3, c_2 = 0.2, \mu = 1$)	155
76	Throughput versus Time Delay (L_1 is fixed, $C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 1.0$)	156
77	Throughput versus Time Delay (L_2 is fixed, $C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 1.0$)	156
78	Queueing Model for Bottleneck Decentralised Flow Control	162
79	Transition 'Up' and 'Down' Rate of GE(n)/GE(n)/c/N Queue with Two Classes of Users	165
80	Throughput versus Control ($C_{\mu}^2 = C_{\lambda_1}^2 =$ $C_{\lambda_2}^2 = 0.5, c_1 = 0.6, c_2 = 1.0, \mu = 1, c = 2$)	197
81	Throughput versus Control ($C_{\mu}^2 = C_{\lambda_1}^2 =$ $C_{\lambda_2}^2 = 1.0, c_1 = 0.6, c_2 = 1.0, \mu = 1, c = 2$)	197
82	Throughput versus Control ($C_{\mu}^2 = C_{\lambda_1}^2 =$ $C_{\lambda_2}^2 = 2.0, c_1 = 0.6, c_2 = 1.0, \mu = 1, c = 2$)	198
83	Time Delay versus Control ($C_{\mu}^2 = C_{\lambda_1}^2 =$ $C_{\lambda_2}^2 = 0.5, c_1 = 0.6, c_2 = 1.0, \mu = 1, c = 2$)	198

84	Time Delay versus Control ($C_{\mu}^2 = C_{\lambda_1}^2 =$ $C_{\lambda_2}^2 = 1.0, c_1 = 0.6, c_2 = 1.0, \mu = 1, c = 2) \dots$	199
85	Time Delay versus Control ($C_{\mu}^2 = C_{\lambda_1}^2 =$ $C_{\lambda_2}^2 = 2.0, c_1 = 0.6, c_2 = 1.0, \mu = 1, c = 2) \dots$	199
86	Throughput versus Control ($C_{\mu}^2 = C_{\lambda_1}^2 =$ $C_{\lambda_2}^2 = 0.5, c_1 = 1.2, c_2 = 1.4, \mu = 1, c = 3) \dots$	200
87	Throughput versus Control ($C_{\mu}^2 = C_{\lambda_1}^2 =$ $C_{\lambda_2}^2 = 1.0, c_1 = 1.2, c_2 = 1.4, \mu = 1, c = 3) \dots$	200
88	Throughput versus Control ($C_{\mu}^2 = C_{\lambda_1}^2 =$ $C_{\lambda_2}^2 = 5.0, c_1 = 1.2, c_2 = 1.4, \mu = 1, c = 3) \dots$	201
89	Throughput versus Control ($C_{\mu}^2 = C_{\lambda_1}^2 =$ $C_{\lambda_2}^2 = 5.0, c_1 = 1.2, c_2 = 1.0, \mu = 1, c = 3) \dots$	201
90	Throughput versus Control ($C_{\mu}^2 = C_{\lambda_1}^2 =$ $C_{\lambda_2}^2 = 5.0, c_1 = 1.2, c_2 = 0.6, \mu = 1, c = 3) \dots$	202
91	Time Delay versus Control ($C_{\mu}^2 = C_{\lambda_1}^2 =$ $C_{\lambda_2}^2 = 0.5, c_1 = 1.2, c_2 = 1.4, \mu = 1, c = 3) \dots$	202

92	Time Delay versus Control ($C_{\mu}^2 = C_{\lambda_1}^2 =$ $C_{\lambda_2}^2 = 1.0, c_1 = 1.2, c_2 = 1.4, \mu = 1, c = 3) \dots\dots$	203
93	Time Delay versus Control ($C_{\mu}^2 = C_{\lambda_1}^2 =$ $C_{\lambda_2}^2 = 5.0, c_1 = 1.2, c_2 = 1.4, \mu = 1, c = 3) \dots\dots$	203
94	Time Delay versus Control ($C_{\mu}^2 = C_{\lambda_1}^2 =$ $C_{\lambda_2}^2 = 5.0, c_1 = 1.2, c_2 = 1.0, \mu = 1, c = 3) \dots\dots$	204
95	Time Delay versus Control ($C_{\mu}^2 = C_{\lambda_1}^2 =$ $C_{\lambda_2}^2 = 5.0, c_1 = 1.2, c_2 = 0.6, \mu = 1, c = 3) \dots\dots$	204
96	Throughput versus Time Delay (L_1 is fixed, $C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 1.0, \mu = 1, c = 3) \dots\dots\dots$	205
97	Throughput versus Time Delay (L_2 is fixed, $C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 1.0, \mu = 1, c = 3) \dots\dots\dots$	205

LIST OF ABBREVIATIONS

CPU	:	Central Processing Unit
FCFS	:	First-Come-First-Served
GE	:	Generalised Exponential
H_2	:	Two Phase Hyperexponential
I/O	:	Input/Output
ME	:	Maximum Entropy
N-ME	:	Norton-Maximum Entropy
PME	:	Principle of Maximum Entropy
S	:	Simulation

Abstract of dissertation submitted to the Senate of Universiti Pertanian Malaysia in fulfilment of the requirements for the degree of Doctor of Philosophy.

**ANALYSIS AND DECENTRALISED OPTIMAL FLOW CONTROL
OF HETEROGENEOUS COMPUTER COMMUNICATION
NETWORK MODELS**

by

KU RUHANA KU MAHAMUD

December 1993

Chairman : Assoc. Prof. Dr. Abu Talib Othman

Faculty : Science and Environmental Studies

General closed queueing networks are used to model the local flow control in multiclass computer communication networks with single and multiple transmission links. The problem of analysing multiclass general closed queueing network models with single server and multiserver is presented followed by the problem of decentralised optimal local flow control of multiclass general computer communication networks with single and multiple transmission links. The generalised exponential (GE) distributional model with known first two moments has been used to represent general interarrival and transmission time distributions as various users have various traffic characteristics.

A new method of general model reduction using the Norton's theorem for general queueing networks in conjunction with the universal maximum entropy algorithm is proposed for the analysis of



large general closed queueing networks. This extension to Norton's theorem has an advantage over the direct application of the universal maximum entropy approach whereby the study of a subset of queueing centres of interest can be done without repeatedly solving the entire network.

The principle of maximum entropy is used to derive new approximate solutions for the joint queue length distributions of multiclass general queueing network models with single server and multiserver and favourable comparisons with other methods are made. The decentralised optimal local flow control of the multiclass computer communication networks with single and multiple transmission links is shown to be a state dependent window type mechanism that has been traditionally used in practice. The maximum number of packets in transit within the system corresponding to a maximum throughput and can be determined from a preassigned upper bound on the mean time delay, the average allowed load and the parameters of the underlying systems. The direct dependence of the maximum throughput on the mean time delay is also determined. The optimal local flow control with global objectives results in a team decision that does not favour any individual user, and depends only on the relative order of their packet generation rates.

Numerical examples provide useful informations on how critically system behaviour is affected by (i) the distributional form of the interarrival and transmission patterns, (ii) the maximum input rate. The analytic results described in this thesis suggest that (i) analytical analysis for general closed queueing networks which are

used to model computer communication networks can be analysed using the principle of maximum entropy, (ii) congestion problems in computer communication networks with non-exponential data flows should be examined in terms of maximum throughput under a time delay constraint where the offered load appears only as a parameter.

Abstrak disertasi yang dikemukakan kepada Senat Universiti Pertanian Malaysia bagi memenuhi syarat untuk ijazah Doktor Falsafah

**ANALISIS DAN KAWALAN ALIRAN OPTIMA NYAHPUSAT
BAGI MODEL-MODEL
RANGKAIAN KOMUNIKASI KOMPUTER HETEROGEN**

Oleh

KU RUHANA KU MAHAMUD

Disember 1993

Pengerusi : Prof. Madya Dr. Abu Talib Othman

Fakulti : Sains dan Pengajian Alam Sekitar

Rangkaian-rangkaian giliran umum tertutup digunakan untuk memodel kawalan aliran setempat dalam rangkaian komunikasi komputer multi-kelas dan satu atau beberapa pautan transmisi. Masalah bagi menganalisis model-model rangkaian giliran tertutup multi-kelas dengan satu-pelayan dan multi-pelayan diperkenalkan diikuti oleh masalah kawalan aliran optima nyahpusat setempat untuk rangkaian-rangkaian komunikasi komputer umum multi-kelas dengan satu dan beberapa pautan transmisi. Model taburan eksponen teritlak (GE) dengan mengetahui dua momen pertama digunakan untuk mewakili taburan-taburan masa antara ketibaan dan transmisi di mana setiap pengguna mempunyai ciri-ciri trafik yang berbagai.

Satu kaedah baru penurunan model umum menggunakan teorem Norton untuk rangkaian giliran umum dan algoritma entropi maksimum umum dicadangkan bagi menganalisis rangkaian-rangkaian giliran tertutup yang besar. Lanjutan kepada teorem Norton ini mempunyai satu kelebihan

dari penggunaan terus entropi maksimum umum yang mana kajian berparameter bagi satu subset pusat giliran yang diminati boleh dilakukan tanpa mengulang menyelesaikan keseluruhan rangkaian.

Prinsip entropi maksimum diguna untuk menerbitkan penyelesaian anggaran baru taburan-taburan gabungan panjang giliran untuk rangkaian giliran umum tertutup multi-kelas dengan satu-pelayan dan multi-pelayan dan perbandingan yang baik dengan kaedah-kaedah lain dibuat. Kawalan aliran optima nyahpusat setempat untuk rangkaian komunikasi komputer multi-kelas dengan satu atau beberapa pautan transmisi adalah mekanisma jenis tingkap bersandar kepada keadaan yang digunakan secara tradisi. Bilangan maksimum paket dalam transit sistem berpadanan dengan perolehan maksimum dan boleh ditentukan dari batas atas yang diberi untuk min masa lewat, purata beban dan parameter-parameter sistem. Persandaran terus perolehan maksimum ke atas min masa lewat juga ditentukan. Kawalan aliran optima setempat dengan objektif sejagat menghasilkan keputusan kumpulan yang tidak mengambil kira setiap pengguna dan sebaliknya bergantung pada giliran relatif pengeluaran paket.

Contoh-contoh numerik memberi maklumat penting tentang bagaimana kritikal perilaku sistem dipengaruhi oleh (i) bentuk taburan antara ketibaan dan corak transmisi (ii) kadar maksimum input. Keputusan analitik dalam tesis ini mencadangkan (i) analisis analitik untuk rangkaian giliran tertutup umum yang digunakan untuk memodel rangkaian-rangkaian komunikasi komputer boleh dianalisis dengan menggunakan prinsip entropi maksimum (ii) masalah kesesakan dalam rangkaian komunikasi komputer dengan aliran data bukan eksponen patut

dikaji dalam sebutan perolehan maksimum di bawah kekangan masa lewat di mana beban yang diberi merupakan satu parameter sahaja.

CHAPTER I

INTRODUCTION

The interest in computer communication networks has grown enormously in the last thirty years. The growth in the need for data communications comes from a large number of varied application areas. The finance industry including banking and insurance firms, has a growing need for remote data processing (i.e. electronic fund transfers etc.). In the field of medicine and health, there is a need for large information banks with remote access. Educational computing needs currently emphasizes interactive use as opposed to routine data entry, retrieval and acquisition. Large government agencies have vast data exchange requirements. Tactical military computer communication needs are perhaps the largest and most demanding (i.e. secure, rapid, reliable). Application such as these and many others that are providing the manpower, time and money behind the enormous growth of the information processing industry. As the result, computer systems and computer communication networks become highly complicated and complex in term of data management where different classes of users have their own characteristic and quality of service (QOS) requirements such as throughput, time delay, cost and error rate. In this context, a computer network can be viewed as a facility that makes possible communication between computers and other devices. Major components of the network include the connecting links, the interface between devices and the network, and protocols, which are rules for managing the network resources.

Computer systems and computer communication networks can be modelled as queueing networks whose centres represent resources (i.e. processors, memory and communication devices) and whose branches represent paths of information flow. Queueing network models have been used to evaluate the performance of computer systems and computer communication networks since the early 70's (Kleinrock, 1975b; Allen, 1978). As a result, the analysis of the queueing networks has received an enormous impetus.

Communication to and from digital computers as shown in Figure 1 is often characterised by the flow of large numbers of very short messages with high coefficient of variation which is typical of most interactive data processing applications (Lam, 1978; Pawlita, 1981). The behaviour of users accessing computing power is extremely bursty; that is the typical user may be characterised as one who makes demands rarely but when he does he requires a high bandwidth in terms of communication as well as in terms of processing. Therefore in modelling heterogeneous computer communication systems, it is more realistic to consider the second moments of the distribution concerned which leads to the use of general queueing networks.

This thesis is concerned with the analyses of single server and multiserver general closed queueing network models with multiclass users which is of interest to computer systems and computer communication networks. The thesis is also concerned with the analyses of the optimal flow control of multiclass computer communication networks with single and multiple transmission links represented by the general closed queueing network models.

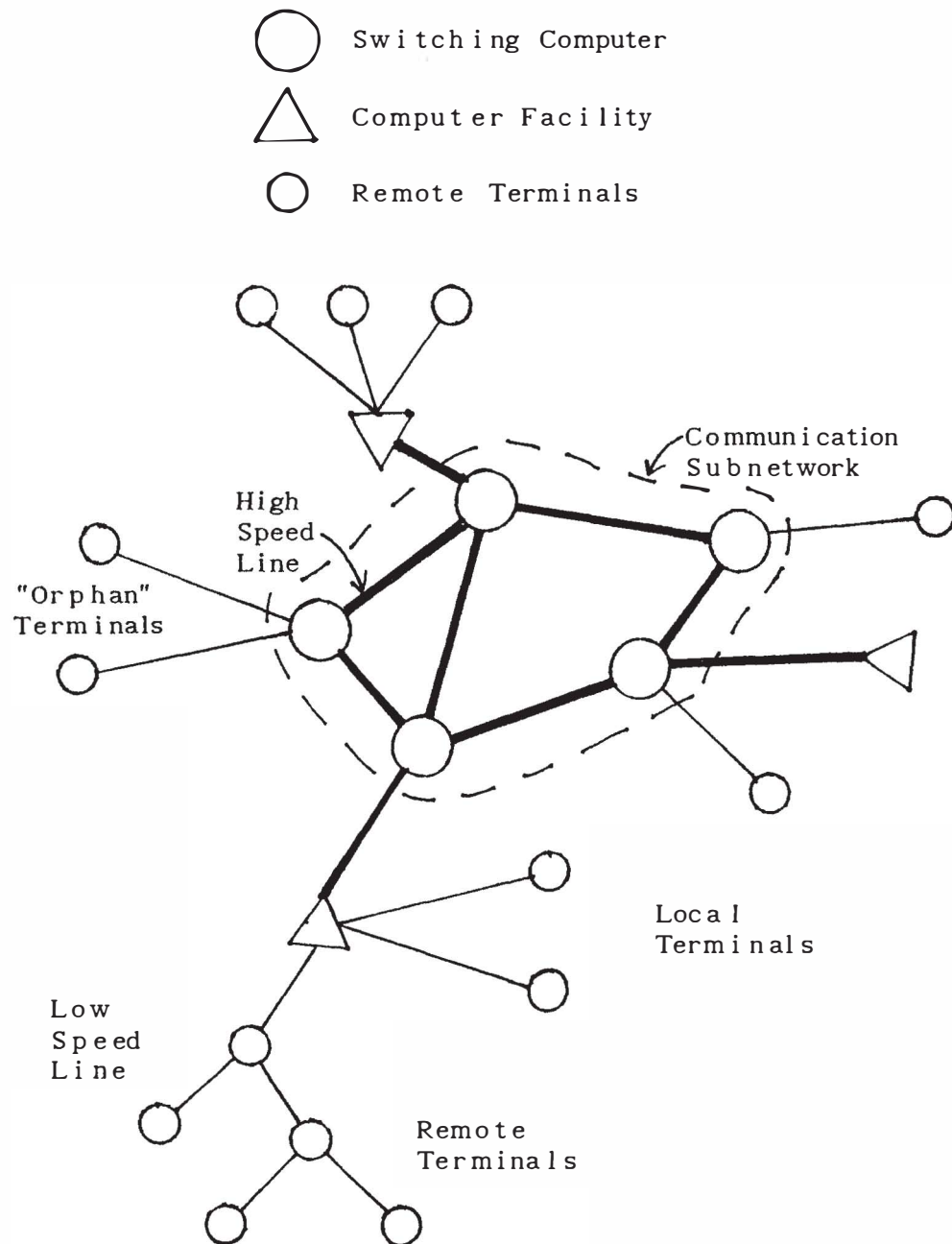


Figure 1 : The Structure of a Computer Communication Network

In the context of general queueing network models, the generalised exponential (GE) distributional model of the form (El-Affendi and Kouvatsos, 1983; Kouvatsos, 1986a; 1988a)

$$f(t) = \left[\frac{C^2 - 1}{C^2 + 1} \right] u_0(t) + \frac{4\mu}{(C^2 + 1)^2} \exp \left[\frac{-2\mu t}{C^2 + 1} \right], \quad t \geq 0, \quad [1.1]$$

where μ is the mean service rate, C is the coefficient of variation and $u_0(t)$ is the unit impulse function, has been used to approximate general interarrival and service time distribution with known first two moments. This model is robust and versatile due to its memoryless properties and has been shown (Kouvatsos, 1988a) to maximise the entropy function subject to mean value constraints based on μ and C (see Appendix A). Furthermore, it facilitates the derivation of product form approximations for isolated queues and networks (Kouvatsos, 1986a).

In the remaining of this chapter we introduce the background of the problems associated with general closed queueing network models used in the performance analysis of computer systems and computer communication networks. We also present the objectives of the study which addresses the analysis of single server and multiserver general closed queueing networks with multiclass users and the use of the queueing network models in determining the optimal flow control of computer communication networks. This chapter also highlights the needs to analyse the general closed queueing network models. Lastly we present a summary of the thesis.

Background of the Problem

Many of the issues involved in the consideration of computer systems and computer communication networks deal with the allocation of resource among competitive demands for such resources. These demands may rise, for example, from the huge number of terminals which are accessing resources of various kinds in the form of processing capacity, storage capacity and transmission capacity. A means for allocating these resources in order to resolve the conflicting demands is one of the most important aspects of today's system design and operation. These conflicting demands naturally give rise to queues and delays to different kind of users. The application of queueing theory is very much required in assisting the understanding of the network behaviour. In this case, the computer communication network can be represented by a general queueing network model with multiclass users.

There has been increasing interest in performance analysis of parallel processing on multiprocessor systems composed of a finite number of homogeneous processors (Gelenbe and Lou, 1988; Bambos and Walrand, 1991). Performance predictions for concurrent programs on multiprocessor systems are of crucial importance for both software and hardware designers. The multiprocessor computer system can also be represented by a general queueing network with multiserver centres and multiclass users.

Effective computer communication network configurations are also important in the areas of safety and strategic applications such as on line credit card validation systems, military applications and computer integrated manufacturing which are currently in great demand. In order to achieve this the computer communication networks will

employ a built in redundancy (i.e. critical portions of the equipment being duplicated and cross connected). Hence there are possibilities that a point-to-point communication between node-to-node will consist of multiple links. The analysis of multiserver general queueing network models with multiserver centres which represent the multiple transmission links, will be of great use to the computer communication networks.

There are situations in a network, where the offered traffic exceeds the load that can be efficiently handled by the network. The network is then unable to meet the quality of service requirements of the various classes of users, and this phenomenon is characterised by the term congestion. Congestion control is one of the main issue considered in designing a computer communication network. A typical example of congestion occurrence is shown in Figures 2 and 3, which depict general patterns of throughput and time delay in a network, as the load of the offered traffic increases. In the low load region, throughput increases as the offered load increases. As the load approaches the network capacity, throughput increases very slow, until it stops increasing, when the network capacity is reached. If the offered load is increased beyond this point, queue lengths build up and eventually the nodal buffers become full, causing arriving packets to get dropped. Dropped packets will be retransmitted, leading to a waste in network resources. In Figure 2, this is shown as a sudden drop of the throughput-load curve. The time delay curve follows a similar pattern where time delay increases slowly under light loading conditions, faster under moderate load, and eventually increases drastically as nodal buffers start to overflow.

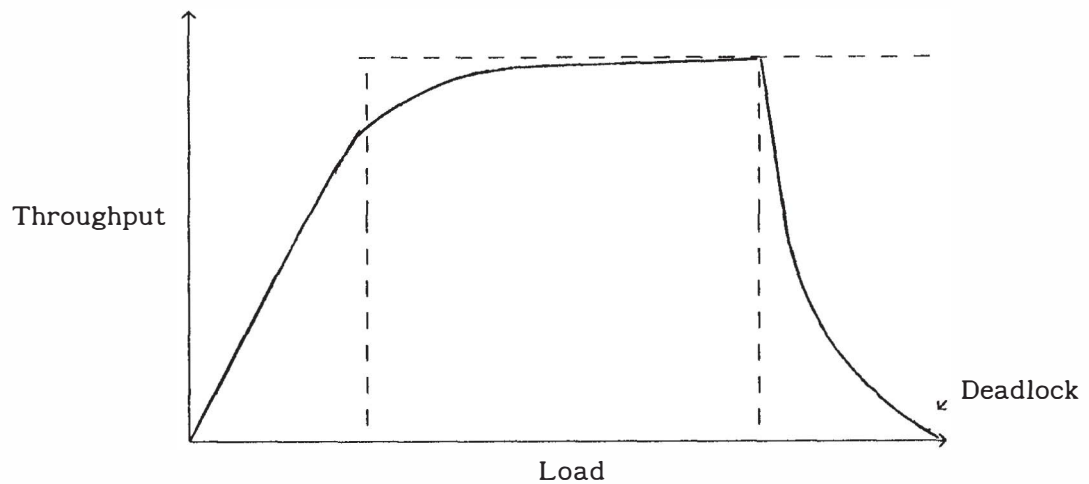


Figure 2 : Typical Plot of Throughput versus Offered Load

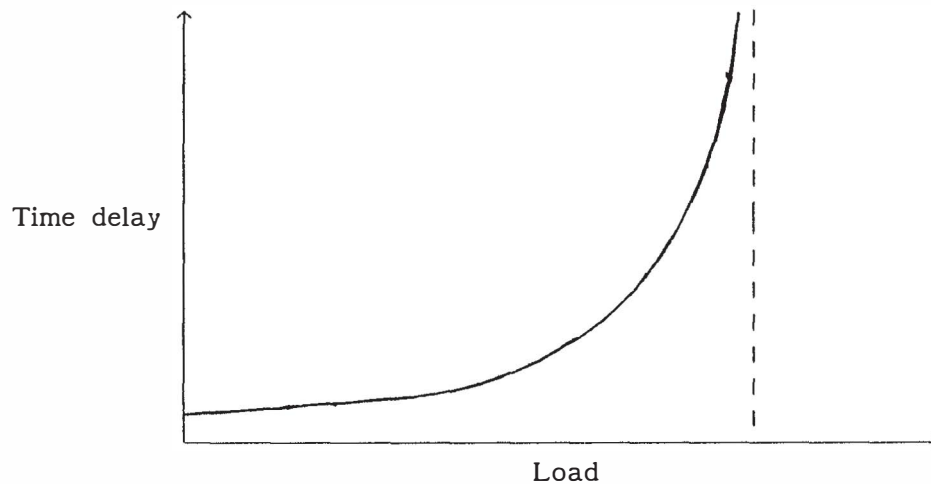


Figure 3 : Typical Plot of Time Delay versus Offered Load

With the rapid growth of computer communication networks control and management, the network flow control has evolved as a distinct class of algorithms and substantial research effort has been devoted to its study. Flow control has been used to deal with this problem where input traffic regulation is performed by the network users themselves. Typically, the users determine the load of traffic they offer to the network based on some feedback information about the

network state such as the level of congestion. As already stated, flow control aims to alleviate and/or prevent congestion, defined as the load of input traffic exceeding its capacity. The high level goal of flow control can be defined as the regulation of input traffic in a way that maintains an efficient trade-off between high network utilisation and quality of service. Typically, network utilisation is quantified in terms of average throughput, and quality of service by average time delay. In a single class network (Kleinrock, 1979; Kouvatsos and Othman, 1989a; 1989b; 1989c; 1989d; Lazar, 1983a; 1983b; 1984; 1991), and in a multiclass user network (Hsiao and Lazar, 1984a; 1984b) the goal of flow control has been defined as maintaining an efficient trade-off between throughput and average time delay. Performance measures have been defined in an effort to quantify this trade-off and compare various flow control schemes. Various techniques have been described in the literature which in general can be classified into three types known as local flow control (Pennotti and Schwartz, 1975), end-to-end flow control (Pennotti and Schwartz, 1975) and global flow control (Davies, 1972). The local and end-to-end flow controls involve restrictions placed on messages of each user while global flow control is concerned with limiting the total number of all messages in the entire network.

Under the local flow control technique, decisions are made at each node whether to transmit or to hold a particular user's message based on the number of messages at the next node on its route. In the end-to-end control technique, the total number of messages allowed from one source to one destination is restricted at any one time. The global flow control technique has been termed 'isarithmic' by Davies (1972) and involves limiting the total number of all users' messages

in the entire network. This control technique is designed to operate in addition to either one of the earlier mentioned technique. The next step taken was to seek for control strategies that optimise the network performance with respect to those performance measures, giving rise to a special class of problems which is called optimal flow control. The choice of the optimisation criterion depends on the particular application. In practice, the average throughput and the average time delay are two quantities of interest in considering the performance of a computer communication network and its underlying protocols. In this thesis, the local flow control will be used to analyse the computer communication networks with multiclass users with single and/or multiple transmission links.

Objectives of the Study

The main objective of the research is to address the analysis of general closed queueing networks that are used to model the computer systems and computer communication networks. The prime aims are:

- i. To derive new approximate analytic results for single server and multiserver general closed queueing network models with multiclass users of first-come-first-served (FCFS) service discipline and load dependent interarrival and/or service times.
- ii. To determine analytically the optimal flow control of multiclass computer communication networks with single and multiple transmission links represented by general queueing network models.

Significance of the Study

Computer systems and computer communication networks are examples of systems which are sufficiently complex and carefully developed models are necessary for understanding system performance. This study focuses on the analysis of general closed queueing network models which is of great use in the performance analysis of computer systems and computer communication networks. The study also focuses on the optimal flow control of computer communication networks with multiclass users and single and/or multiple transmission links.

The study of flow control has so far focused on single and/or multiclass users with single transmission link and exponential distributed transmission time (Kleinrock, 1975b; Lazar, 1983a; Hsiao and Lazar, 1990; 1991) except for Kouvatsos and Othman (1989a) which focuses on single class users and single transmission link with general transmission time. The study of flow control also focuses on single class users and multiple transmission links (Lazar, 1984; Kouvatsos and Othman, 1989b). The lack of flow control study on computer communication networks with multiclass users and multiple transmission links is due to the fact that very few analytic results were available for closed queueing networks with multiclass users and multiserver centres. This is why the research is focused on such queueing networks.

Traditionally queueing theory has been dealing with single server and multiserver exponential queueing network models (Baskett *et al.*, 1975; Allen, 1978). These models are inadequate to be used in the performance analysis of parallel processing of multiprocessors computer systems. This research is an effort to study more general

class of queueing network models that is associated with parallel processing and multiprogrammed computer systems. The study in this case will focus on single server and multiserver general queueing network models with multiclass users.

In today's world of high technology, computer communication networks are in use in many places. These systems are expensive to design. Once they are installed, it is costly to improve their efficiency. When designing a system, it is often difficult to decide which of the many possible alternatives would give the best performance. After a system is running, it is a complicated task to improve its operation and plan for future changes. If one ends up with unacceptable performance with a reasonable amount of hardware, then the only options are to abandon the system entirely or to go through redesign and redevelopment phases until the system is acceptable. Therefore, performance analysis of single server and multiserver general queueing network models with multiclass users which represent the computer systems and computer communication networks is important.

The analysis of the general closed queueing networks which have to be carried out was based on the principle of maximum entropy (Jaynes, 1957) to produce the maximum entropy queue length distributions of the queueing models. Earlier work by El-Affendi and Kouvatsos (1983) and Kouvatsos (1984; 1986a; 1988a; 1988b) has applied the principle of maximum entropy and its generalisation to derive new analytical results for arbitrary FCFS networks with class switching and the number of users for each class is not fixed. The numerical accuracy of maximum entropy results have so far proved to be very compatible with those obtained from simulation models and relatively

superior to other approximate solutions examined (El-Affendi and Kouvatsos, 1983; Kouvatsos, 1986a; 1988a; Wu and Chan, 1989).

Simulation results, when exact results are not available will provide a sound basis for making relative comparisons on the numerical accuracy of the above solution techniques with maximum entropy method. A simulation package called SIMSCRIPT II.5 (Russell, 1989) is used where appropriate.

It is hoped that this research will be useful not only in obtaining analytical results for the general closed queueing network models but used as a flexible tool in operating and upgrading the computer systems and computer communication networks with non exponential data flows.

Organisation of the Thesis

This thesis is organised as follows. Chapter I provides the introduction to the thesis. It highlights the problems associated with general closed queueing network models used in the performance analysis of computer systems and computer communication networks. It also discusses the needs to analyse complex closed queueing network models with multiclass users, single server and multiserver centres. In Chapter II, a review of related literature on single server and multiserver general queueing network models with multiclass users is presented. The principle of maximum entropy (PME) is then addressed and a discussion of model reduction using Norton's theorem for large general queueing networks in conjunction with the principle of maximum entropy is included. Chapter III is devoted to a method of model reduction for large general queueing networks. This method is based on

the maximum entropy algorithm and Norton's theorem. In Chapter IV the approximate queue length distribution is derived by applying the principle of maximum entropy on the single server general closed queueing system with multiclass users of FCFS service discipline. The one-step maximum entropy solutions and an extension to load dependent servers are also presented. The approximate queue length distribution of the multiserver general closed queueing system with multiclass users of FCFS service discipline is derived in Chapter V. Again a one-step maximum entropy solution and an extension to load dependent servers are presented. The results obtained in Chapter IV will be used in Chapter VI to determine the decentralised optimal flow control of computer communication networks with multiclass users and single transmission link. Chapter VII will make use of the results obtained in Chapter V, to determine the decentralised optimal flow control of computer communication networks with multiclass users and multiple transmission links. Finally, Chapter VIII rounds up the study with a summary of the main conclusions and comments on the prospects of future research.

CHAPTER II

MAXIMUM ENTROPY OF GENERAL QUEUEING NETWORK MODELS

Introduction

In this chapter, a review of related literature on complex general queueing networks with multiclass users and single and/or multiserver centres is presented. This chapter also analyses the related literature on the performance measures that have been used in the optimal flow control of single users and multiclass users network models. A method of obtaining the probability distribution of the states of a queueing system by maximising the corresponding entropy is also presented followed by a discussion of model reduction by using Norton's theorem for queueing network in conjunction with the principle of maximum entropy. In this context, a new algorithm will be presented for reducing a large structure of a general queueing network model to a cyclic queue with two composite centres.

Review of Related Literature

Closed queueing network models have been widely recognised as powerful tools for analysing discrete flow systems such as large scale computer (Takacs, 1962; Guillon, 1977; Allen, 1978) and communication networks (Kleinrock, 1975b; Schwartz, 1987) and constitute a very important tool of investigation into computer systems performance, whether for predicting the behaviour of new designs or proposed changes to existing systems (Lazar, 1982a; 1982b; Kouvatsos and Othman, 1989a; 1989b; 1989c; 1989d; Hsiao and Lazar, 1990; 1991).

However, despite the advances made in the analysis of these models, little is known about the behaviour of single server and/or multiserver general closed queueing networks with multiclass users. Most queueing network models either make gross assumptions to assure an exact numerical solution or employ an approximate method to solve a more general (Kobayashi, 1974; Chandy *et al.*, 1975b; Sauer and Chandy, 1975; Reiser and Lavenberg, 1980) and therefore a more realistic model. When the credibility of the model is severely affected, simulation or other approaches are used in spite of their relatively high cost (Iglehart and Shedler, 1978; Pawlikowski, 1990).

Single Server General Queueing Networks with Multiclass Users

Analytically, little is known about the behaviour of single server general queueing network with multiclass users except in some restricted cases for example when the service discipline is FCFS and the service time are exponentially distributed (Baskett *et al.*, 1975; Reiser and Kobayashi, 1974a; 1974b). Baskett *et al.* (1975) produced one of the classical results on single server general queueing networks with multiclass users where they have succeeded in obtaining the joint equilibrium distribution of queue sizes. Later Reiser and Kobayashi (1974b) presented a generalised results of Baskett *et al.* (1975) to the case of several closed subchains and a constructive derivation of the product form solution.

Kouvatsos (1985) has used the principle of maximum entropy in deriving a product form solution for general open and closed queueing networks with FCFS service discipline and multiclass users under class

switching. Later Kouvatsos (1989) characterised a product form approximation for arbitrary open networks of queues at equilibrium with infinite capacities, single servers, multiclass users, distinct general interarrival and service time distribution per class, non-priority service disciplines and random routing under class switching. The approximation suggests a decomposition of the open network into individual multiclass $G/G/1$ queues with a revised arrival process for each class of users.

More recently, Kouvatsos and Tabet-Aouel (1990), based on the principle of maximum entropy, obtained a product form approximation for general closed queueing networks with a mix of priority and non-priority servers. Their approximation can be interpreted as the truncated state space solution of a decomposition of the network into individual $G/G/1$ and $G/G/\infty$ queues with multiclass users satisfying constraints on population and flow conservation. Earlier on, Kouvatsos and Tabet-Aouel (1989) derived closed form approximations for the marginal queue length distribution per priority class of a stable $G/G/1$ queue and established new one-step recursions for the maximum entropy states probabilities. The principle of maximum entropy is used under two different sets of mean value constraints. This result is then extended by Kouvatsos (1989) in his maximum entropy analysis of general open networks with multiclass users and mixed service disciplines. The maximum entropy approximation suggests a decomposition of the open network into individual multiclass $G/G/1$ queues at equilibrium with a revised arrival process for each class of users.

Multiserver General Queueing Networks with Multiclass Users

A stable multiserver queue with generally distributed interarrival and service times, ($G/G/c$, $c \geq 2$, multiclass users) and finite number of users is very important for the study of computer system and packet switch networks with parallel processors or transmission links. Unfortunately, due to the complexity involved, very few exact or approximate results can be found in the literature (Buzen and Bondi, 1983; Tabet-Aouel and Kouvatsos, 1992). Buzen and Bondi (1983) derived approximations for the mean response times of a stable $M/M/c/PR$ ($c \geq 2$) queue with an arbitrary number of user classes. These results are computationally stable and easily evaluated. Moreover they turn out to be exact when all priority classes have the same mean service times. The analysis is based on a heuristic relationship expressing the mean response times of a $M/M/c/PR$ ($c \geq 2$) queue in terms of those $M/M/c/FCFS$ ($c \geq 1$, single class user) and $M/M/1/PR$ queues.

Recently, Tabet-Aouel and Kouvatsos (1992) proposed an analytic approximations for the mean response times of $R (\geq 2)$ priority classes in a stable $G/G/c/PR$ queue with $c (\geq 2)$ parallel servers under preemptive resume (PR) scheduling. The analysis is based on the extension known heuristic arguments and earlier results regarding the study of the stable $GE/GE/c/FCFS$ ($c \geq 1$, single class) and $GE/GE/1/PR$ queues. Comments on the role of the new mean response time expressions towards the approximation of the joint and marginal queue length distributions of a stable $G/G/c/PR$ queue were also included.

Most of the results obtained for multiserver queue are those with single class users with exponential interarrival or service time distribution (De Smit, 1973a; 1973b; Kelton and Law, 1985; Kimura, 1991a; Knessl *et al.*, 1990; Woodside *et al.*, 1984; Yao, 1985). Other results such as Hokstad (1984), Kimura (1991b; 1991c; 1992), Pujolle and Soula (1979), Shore (1988a; 1988b) and Takacs (1962) analysed single class, multiserver queue with independent and identically distributed interarrival and service times. Kouvatsos and Almond (1988b) has derived new one-step recursions for the queue length distribution of a two centres cyclic queueing network with multiple general servers and a fixed number of users.

Large General Queueing Networks with Single Class Users, Single Server and/or Multiserver Centres

Large general queueing networks under FCFS service discipline do not possess product form solutions and are difficult to analyse. Stochastic analysis (i.e. exact queueing theory) used by Buzen (1973) has produced an efficient computational algorithm for closed exponential networks, based on a simple recursion formula for calculating performance metrics such as throughput and response time at each queueing centre. The product form solution was extended by Baskett *et al.* (1975) to include multiclass user networks. The computational algorithms were improved for this case by Reiser and Kobayashi (1975), Reiser and Lavenberg (1980), Harrison (1985) and Gordon (1990). Another alternative is to use approximate solution techniques such as diffusion (Reiser and Kobayashi, 1974; Gelenbe and Pujolle, 1976) and maximum entropy (El-Affendi and Kouvatsos, 1983;

Kouvatsos, 1986b; 1989b; 1990; Wu and Chan, 1989). The diffusion approximation has been successful in applications to large open networks with heavy traffic, but this has not generally been the case in the analysis of large closed networks. Hierarchical decomposition using Norton's reduction flow equivalent method (Smith, 1949; Chandy *et al.*, 1975a; 1975b) has been proposed for parametric analysis of some large queueing networks that satisfy local balance. Aggregation (decomposition) has also been proposed by Courtios (1977) and by Sauer and Chandy (1975). Kouvatsos (1986b), Sauer and Chandy (1975) and Sevcik *et al.* (1977) have provided approximate formula to calculate coefficients of variation for the flow processes of aggregated large queueing network subsystems.

Flow Control in Computer Communication Networks

In this section, the problem of optimal flow control of single class and multiclass computer communication network models will be discussed. One of the first performance measures that has been studied in the flow control literature is the generalised power (Kleinrock, 1979) where the throughput time delay trade-off has been chosen as a performance criterion for optimal control of simple queueing systems that have been used as network models. In investigating this trade-off Kleinrock (1979) and Barath-Kummar (1980) have chosen the power as an optimal criterion. This criterion, however, does not display the direct dependence between the maximum throughput and the average time delay. Although the power as a performance measures behaves rather well on the technical level leading to unconstrained optimisation problems, it demonstrates certain weaknesses on an intuitive level.

The main drawback of the power criterion is that it cannot capture explicitly the sensitivity of the carried traffic to time delay. To overcome this, Schwartz (1980) suggested the problem of finding a state dependent control to a queueing system with an exponential server.

Maximisation of throughput subject to a quality of service constraint defined as a maximum acceptable average time delay has been proposed by Lazar (1983a) as an appropriate optimisation objective for a data link layer flow control scheme with time outs. Lazar (1983a) has proved that the optimal local flow control of a single class M/M/1 queue is of window type, which, for an offered load maximises the throughput under the constraint of the average time delay not exceeding a pre-assigned value. The window type control mechanism has also been studied by authors such as Davies (1972), Lam and Reiser (1979) and Pennotti and Schwartz (1979). Kouvatsos and Othman (1989a) has extended the results obtained by Lazar (1983a) to include the G/G/1 queue. Since then the optimal flow control studies which focused on window flow type have been reported (Lazar, 1982a; 1983b; 1984; Kouvatsos and Othman, 1989b; 1989c; 1989d; Robertazzi and Lazar, 1985).

In an integrated environment several users share the network resources each having its own performance objectives and quality of service requirements. The nature of flow control is actually revealed in a multiclass setting, under which the complex interaction among various users with different and often conflicting objectives, have to be studied. Lazar (1982b) and Hsiao and Lazar (1990) have investigated

the optimal flow control problem of multiclass queueing networks with partial information whereby only some information about interfering traffic sharing the same resource is known. The resulting optimal control of the traffic under study is shown to be a window flow control mechanism and the window size is a function of the maximum tolerated time delay, the input capacity and the interfering traffic flow.

Hsiao and Lazar (1984) has also studied the decentralised optimal local flow control of a multiclass M/M/1 queue with global objectives. In their investigation it is shown that the optimal control that maximises the average throughput subject to a bounded time delay constraint is a window type mechanism and the window sizes can be obtained from the throughput time delay function. It has also been demonstrated that the optimal control under the global objective does not favour any individual user and depend only on the relative order of their packet generation rates. Later Hsiao and Lazar (1985) studied the decentralised optimal local flow control problem of a multiclass M/M/1 queue which focused on individual objectives. The solution of the decentralised control problem using the Nash equilibrium has been shown to be a window type mechanism.

Maximum Entropy in Queueing Theory

The main part of the basic results in queueing theory is obtained by supposing that the queueing system is in a steady state condition. Generally, a queueing system has reached a steady state condition if the state of the system becomes essentially independent of the initial

uniquely correct method of inference for estimating probability distributions given information in the form of known mean values. The method can be applied in the performance analysis of general queueing systems since expected values of various distributions of interest such as queue length (Kobayashi, 1974; Reiser and Kobayashi, 1974a) are usually known in terms of the first two moments of the interarrival and service times distributions, which are generally obtained via system measurements.

By using this method of statistical inference, we shall propose solutions for single server general closed queueing system with multiclass users based on the known solution of the single server general queueing system with single class users which has been solved by Kouvatsos (1988a). This will be presented in Chapter IV. Subsequently, the results will be used as the building block for approximating a multiserver general closed queueing system with multiclass users which will be presented in Chapter V. The solutions obtained will include both servers and arrivals with load dependent rates. The new solutions will be used in chapter VI and chapter VII to facilitate the derivation of the decentralised optimal local flow control of multiclass computer communication networks with single and multiple transmission links.

The Principle of Maximum Entropy

Let a system Q has a set X of possible discrete states $\{x_0, x_1, x_2, \dots\}$ which may be finite or countable infinite, and state x_n ($n = 0, 1, 2, \dots$) may be specified arbitrarily. Suppose the available

where $\{\gamma_k\}$, are the Langrangian multipliers determined from the set of constraints [2.3], and Z is given by

$$Z = \exp \{\gamma_0\} = \sum_{\mathbf{x}_n \in X} \exp \left\{ - \sum_{k=1}^m \gamma_k h_k(\mathbf{x}_n) \right\}, \quad [2.5]$$

where γ_0 is a Lagrangian multiplier corresponding to the normalisation constraints [2.2]. The Lagrangian multipliers $\{\gamma_k\}$ satisfy the following relations:

$$- \frac{\delta \gamma_0}{\delta \gamma_k} = \langle h_k \rangle, \quad k = 1, 2, \dots, m, \quad [2.6]$$

and the maximum entropy is expressed by

$$\max_p H(p) = \gamma_0 + \sum_{k=1}^m \gamma_k \langle h_k \rangle. \quad [2.7]$$

Generally, there is no analytical expression for the solution of [2.6]. Fortunately in the queueing theory context, we have simple expressions for the probability distribution satisfying the principle of maximum entropy, subject to the usual available data on the queueing system.

The general solution for the single server general closed queueing network model with multiclass users will be presented in Chapter IV while the general solution for the multiserver general closed queueing network model with multiclass users will be presented in Chapter V.

Norton's Theorem in Queueing Networks

In this section we will describe how Norton's theorem can be applied to solve large general queueing networks whereby the subnetwork under parametric study consists of several centres. In the electrical circuit theory, Norton's theorem (Smith 1949) allows replacement of all but one component of a network by a current source and a parallel internal resistance. The objective is to study the one component without directly considering the others. Hence the analysis of the equivalent circuit requires less computation than the analysis of the original circuit. Chandy *et al.* (1975a) determined a relationship that exists between some queueing networks and electrical networks, with users and throughput (i.e. rate of flow of users) being analogous to electrical charge and electrical current, respectively. The authors show how to directly apply some methods from electrical circuit theory, in particular Norton's theorem to queueing networks that satisfy local balance. We will then extend the work done by Chandy *et al.* (1975a), (i.e. from exponential queueing network to general queueing network) thus generalising the classic Norton's theorem. The approximate solution of general queueing network based on the principle of maximum entropy has been shown by Kouvatsos (1986b) to be of a product form. A method of model reduction using the maximum entropy algorithm in conjunction with Norton's theorem for queueing networks is used for the analysis of general distributed closed queueing networks. This method is called the Norton-Maximum Entropy (N-ME).

Consider a closed queueing network with N users as depicted in Figure 4. Chandy *et al.* (1975a) study the behaviour of subsystem 1 as the subsystem parameters are varied. They constructed an equivalent network in which all the centres except those in subsystem 1 are replaced by a single composite centre (see Figure 5). Let $\chi_2(n)$ be the service rate for the composite centre 2 when n is the number of users waiting for or being served at this centre ($n = 0, 1, \dots, N$) and $\hat{C}_2$ be the composite coefficient of variation. Set $\chi_2(n)$ equal to the rate at which users pass (throughput) through subsystem 1 when there are n

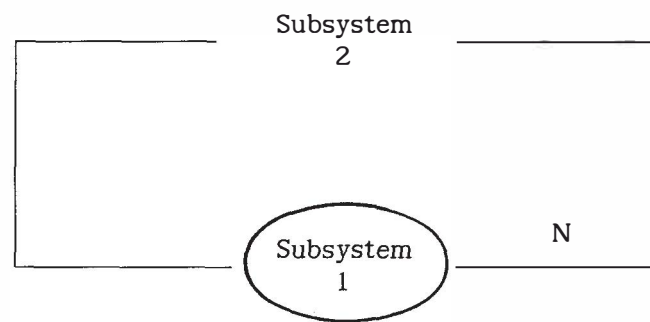


Figure 4 : Queueing Network

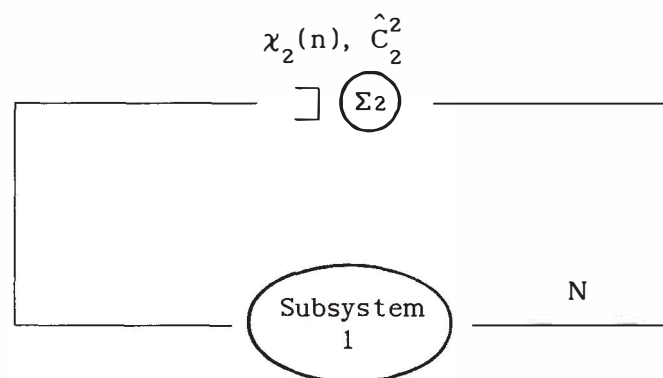


Figure 5 : Equivalent Queueing Network

users in the given network and when subsystem 1 is replaced by a 'short'. This is when the service time of all the centres in subsystem 1 are set to zero (see Figure 6). The behaviour of the subsystem 1 in the equivalent network is the same as in the original network (Chandy *et al.*, 1975a). Since the theorem proposed by Chandy *et al.* (1975a) applies to the entire product form queueing networks, we shall use this approximate and heuristic method to reduce an arbitrary form of network into an amenable form. The analysis of large general queueing networks using the N-ME method will be presented in the following chapter.

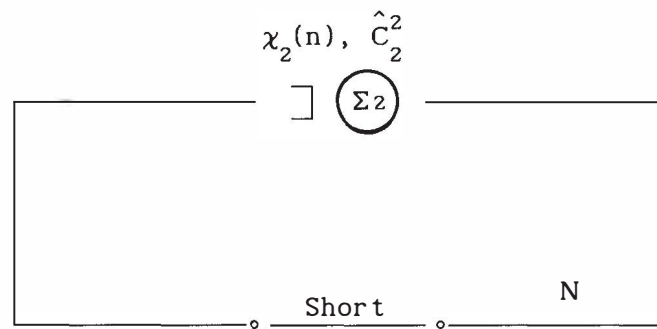


Figure 6 : Subsystem 1 Shorted

Summary

In this chapter, a review on existing literature on complex general queueing network with multiclass users, single and/or multiserver centres has been provided. It can be seen that very few exact or approximate results can be found on such queueing networks. From the review done on flow control, it can be seen that optimal flow

control lays the mathematical framework under which the nature of flow control as a decentralised decision problem can be explicitly modeled and systematically studied. Furthermore, optimal flow control schemes reveal the limits in performance that can be achieved. The principle of maximum entropy has also been described to determine the most random probability distribution subject to a given mean values of random variables. The results are general enough, without any reference to a specific queueing model. A discussion on Norton's theorem for general queueing network has been presented in which a large structure of general queueing network is reduced into a load dependent two stage cyclic queueing network.

CHAPTER III

NORTON-MAXIMUM ENTROPY METHOD FOR LARGE GENERAL QUEUEING NETWORKS

Introduction

In this chapter, an extension of Norton's reduction method to GE type distributed system is proposed. This extension is made possible by the presence of the maximum entropy algorithm for solving general queueing networks which has been shown by Kouvatsos (1986b) to be a universal algorithm for solving large general queueing networks. In this case a large general queueing network can be reduced to an amenable form. This chapter will provide a building block for solving several load dependent centres in the subsequent chapters.

The computer communication networks (i.e. packets or messages are routed instantaneously and sequentially between arbitrary connected resources) can be modelled by large general queueing networks since communication within the networks is often characterised by the flow of large numbers of very short messages with high coefficient of variation (Lam, 1978; Pawlita, 1981). Such models may then be used to study the behaviour of the actual system where certain parameters are varied. Often the parameters being varied are localised to only a part of the network or subnetwork which makes the repeated solution of the entire network unnecessary. Norton's theorem, proved by Chandy *et al.* (1975a) for closed queueing networks which satisfy local balance, states that it is possible to replace all the uninteresting centres by one composite centres with FCFS discipline and an appropriate service rate such that the queue length and waiting time distributions of the

subsystem in the flow equivalent network are identical with the original network. Norton's theorem, however, is proved for the case that subnetwork is a single centre. It can quite easily be extended to cases in which the subsystem has more centres as will be shown later. This concept can be applied to the computation of an approximate solution for large general closed queueing networks for which an exact solution is not known.

The General Queueing Model

Consider a large general closed queueing network of arbitrary interconnected service centres with N circulating users of a single class and M centres as depicted in Figure 7. All centres with unlimited capacity queue provide FCFS service discipline with independent GE type service distribution. Let

- n_i - number of users at centre i , $i = 1, 2, \dots, M$,
- n_j^* = $n_1 + n_2 + \dots + n_j$ for any $j = 1, 2, \dots, M$,
- λ_i - mean arrival rate at centre i , $i = 1, 2, \dots, M$,
- C_{ai}^2 - squared coefficient of variation for interarrival time at centre i , $i = 1, 2, \dots, M$,
- μ_i - mean service rate at centre i , $i = 1, 2, \dots, M$,
- C_{si}^2 - squared coefficient of variation for service time at centre i , $i = 1, 2, \dots, M$,
- X_i - mean throughput rate at centre i , $i = 1, 2, \dots, M$,
- U_i - utilisation at centre i , $i = 1, 2, \dots, M$,
- p_{ij} - routing probability (first order Markov chain) that a user on completing service at centre i proceeds instantaneously to centre j , $i, j = 1, 2, \dots, M$.

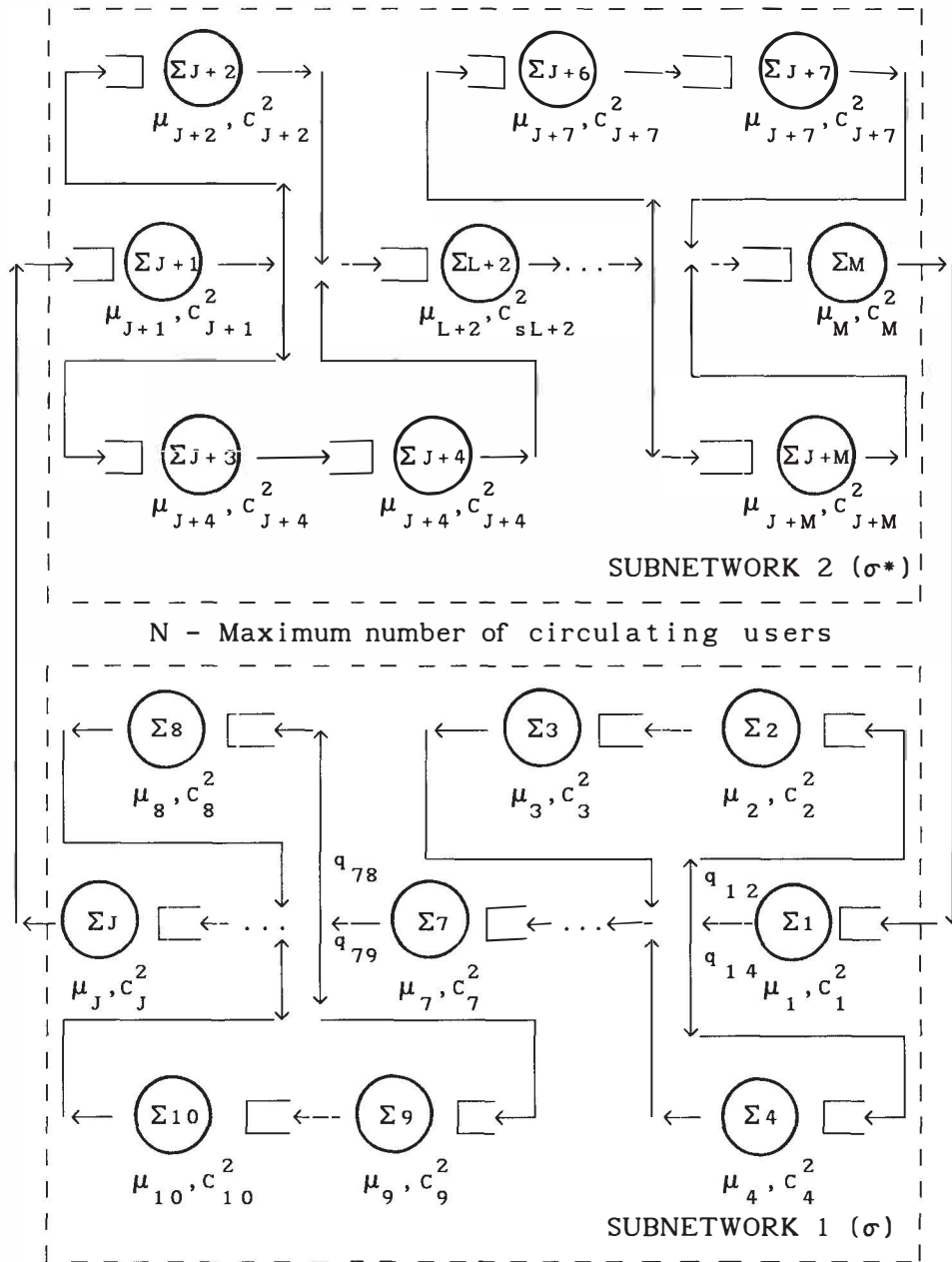


Figure 7 : A General Queueing Network with M Service Centres

Define the state $S = S(N, M)$ of the network as

$$S(N, M) = \{ \underline{n} = (n_1, n_2, \dots, n_M) / \sum_{i=1}^M n_i = N; n_i \geq 0, i=1, 2, \dots, M \} \quad [3.1]$$

It is known that the equilibrium state probability density function for general closed queueing networks using the principle of maximum entropy (Kouvatsos, 1986b) has the form

$$p(\underline{n}) = \frac{1}{Z(N)} \prod_{i=1}^M f_i(n_i), \quad [3.2]$$

where $Z(N)$ is a normalisation constant and $f_i(n_i)$ is defined recursively as follows:

$$f_i(n_i) = \begin{cases} 1 & n_i = 0, \\ g_i x_i & n_i = 1, \\ x_i f_i(n_i - 1) & n_i = 2, 3, \dots, N-1, \\ x_i y_i f_i(n_i - 1) & n_i = N, \end{cases} \quad [3.3]$$

for all $i = 1, 2, \dots, M$, g_i and x_i are the Lagrangian coefficients obtained by solving M corresponding infinite capacity (but stable) GE/GE/1 queues of an open "near equivalent" pseudo network (with no external arrival process and otherwise an identical topology to the original closed network) satisfying the flow balance equations

$$\lambda_i = \sum_{j=1}^M \lambda_j p_{ji}, \quad \text{for } i = 1, 2, \dots, M,$$

and the fixed population mean constraint

$$\sum_{i=1}^M n_i = N.$$

The detailed algorithm to find these coefficients is given in (Kouvatsos, 1986b). The normalising constant $Z(N)$ can be calculated by defining

$$z(n_j^*, j) = \sum_{S(n_j^*, j)} \prod_{i=1}^j f_i(n_i), \quad [3.5]$$

$n_j^* = 0, 1, \dots, N$ ($j = 1, 2, \dots, M$) where

$$S(n_j^*, j) = \{(n_1, n_2, \dots, n_j) / \sum_{i=1}^j n_i = n_j^*; n_i \geq 0, i = 1, 2, \dots, M\}.$$

It can be shown that

$$\begin{aligned} z(n_M^*, M) &= Z(n_M^*), & n_M^* &= 0, 1, \dots, N; \\ z(0, j) &= 1, & j &= 1, 2, \dots, M; \\ z(n_1^*, 1) &= f_1(n_1), & n_1^* &= 0, 1, \dots, N; \\ z(n_j^*, j) &= \sum_{n_j=1}^N f_j(n_j) z(n_{j-1}^*, j-1), & n_j^* &> 0, j > 1. \end{aligned} \quad [3.6]$$

Moreover, by using the maximum entropy recursive form of the queueing station state function $f_j(n_j^*)$, for $n_j^* > 0$, $j > 1$, [3.5] can be written as

$$z(n_j^*, j) = z(n_j^*, j-1) + x_j [(g_j - 1) z(n_j^* - 1, j-1) + z(n_j^* - 1, j)]. \quad [3.7]$$

Finally, since $\sum_{S(N, M)} p(n_1, n_2, \dots, n_M) = 1$, it follows that

$$Z(N) = z(N, M).$$

This type of convolution is repeated iteratively until a third Lagrangian coefficient y_i is determined satisfying the flow balance equation [3.4].

Service Station Aggregation

Renumbering the service centres in the network such that centres 1, 2, ..., J denoted by $\sigma = \{\Sigma_1, \Sigma_2, \dots, \Sigma_J\}$ are in subnetwork 1, and all the remainder centres J+1, J+2, ..., M are in subnetwork 2 denoted by $\sigma^* = \{\Sigma_{J+1}, \Sigma_{J+2}, \dots, \Sigma_M\}$. Subnetwork 1 can be aggregated by "shorted" subnetwork 2 and subnetwork 2 can in turn be aggregated by "shorted" subnetwork 1 (Figures 7-9). Define $p_{\sigma^*}(n_{J+1}, n_{J+2}, \dots, n_M)$ as the marginal probability of a state $(n_{J+1}, n_{J+2}, \dots, n_M)$ of subnetwork 2. Then

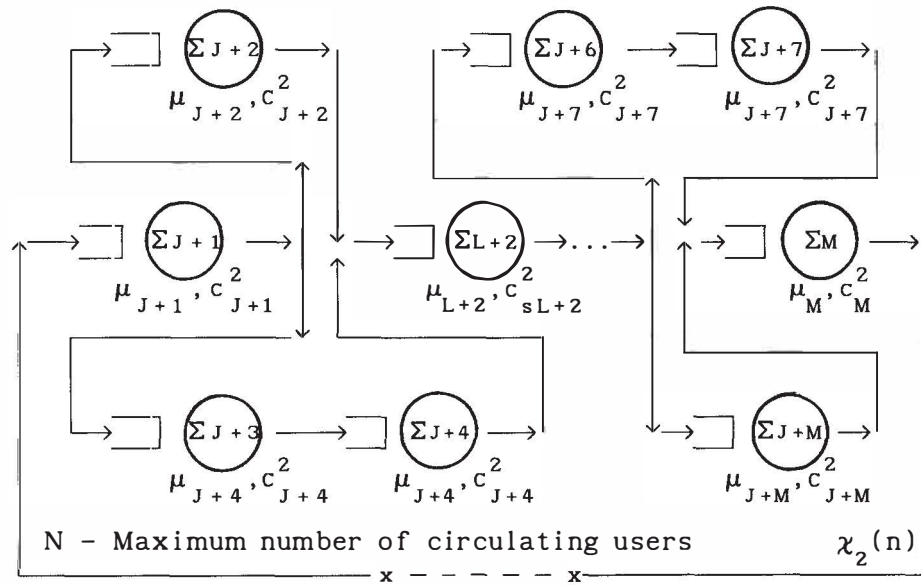


Figure 8 : A "Shorted" Subnetwork 1 (σ) Cyclic Queueing Network

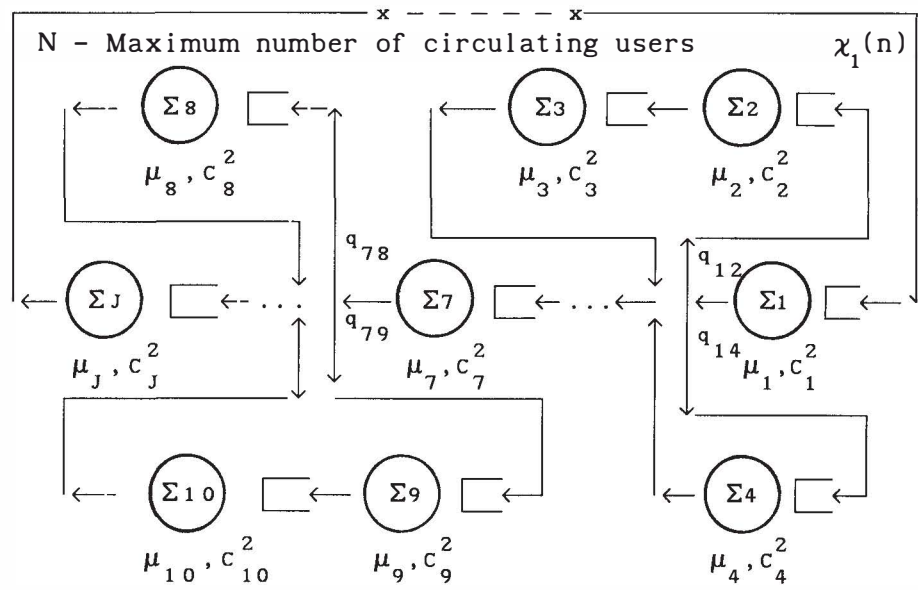


Figure 9 : A "Shorted" Subnetwork 2 (σ^*) Cyclic Queueing Network.

$$p_{\sigma^*}(n_{J+1}, n_{J+2}, \dots, n_M) = \frac{1}{Z(N)} \sum_{(n_1, n_2, \dots, n_J)} \prod_{i=1}^M f_i(n_i), \quad [3.8]$$

where the summation is over all states $(n_1, n_2, \dots, n_J)$ such that $n_1 + n_2 + \dots + n_J = n_J^*$. That is

$$p_{\sigma^*}(n_{J+1}, n_{J+2}, \dots, n_M) = \frac{1}{Z(N)} \prod_{i=J+1}^M f_i(n_i) \sum_{(n_1, n_2, \dots, n_J)} \prod_{i=1}^J f_i(n_i), \quad [3.9]$$

or

$$p_{\sigma^*}(n_{J+1}, n_{J+2}, \dots, n_M) = \frac{1}{Z(N)} z(n_J^*, J) \prod_{i=J+1}^M f_i(n_i). \quad [3.10]$$

This result enables one to perform a parametric analysis of the centres in subnetwork 2 without having to re-evaluate the remainder of the network. In this verification study, subnetworks 1 and 2 are replaced by a flow equivalent to two composite centres C and C^* respectively as depicted in Figure 10. For this purpose, a "shorted"

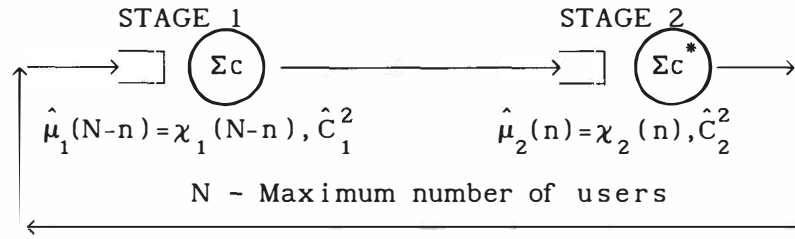


Figure 10 : A Flow Equivalent Load Dependent
Two Stage Cyclic Queueing System

subnetwork 1 cyclic queueing network as depicted in Figure 8 in which the service times of all centres in subnetwork 1 are set to zero (i.e. the mean service rate goes to infinity). Then a "shorted" subnetwork 2 cyclic queueing network is established in which the service times of all centres in subnetwork 2 are subsequently set to zero. Let $\chi_1(n)$, $n = 1, 2, \dots, N$ be the throughput of the shorted cyclic queueing network of Figure 9 calculated by using the universal maximum entropy algorithm (Kouvatsos, 1986b) when there are n users circulating the network. The universal maximum entropy is presented in Appendix B and was also successfully implemented in C programming language. Set the state dependent service rates $\hat{\mu}_1(n)$, $n = 1, 2, \dots, N$ of the composite centre C equal to the rate at which users pass (i.e. throughput) between the shorted subnetwork. Similarly, set the state dependent

service rates $\hat{\mu}_2(n)$, $n = 1, 2, \dots, N$ of the composite centre C^* equal to the rate at which users pass $\chi_2(n)$ between the shorted subnetwork of Figure 8.

The main source of error is in determining the squared coefficients of variation $\hat{C}_1^2$ and $\hat{C}_2^2$ to be used for the composite centres C and C^* . Based on numerical studies with respect to work done by Kouvatsos (1986b), Sauer and Chandy (1975) and Sevcik *et al.* (1977), the proposed squared coefficients of variation are taken to be

$$\begin{aligned}\hat{C}_1^2 &= \begin{cases} C_{s\ i}^2, & \text{if } C_{s\ i}^2 = C_s^2 \text{ for } i = 1, 2, \dots, J, \\ 1, & \text{otherwise,} \end{cases} \\ \hat{C}_2^2 &= \begin{cases} C_s^2, & \text{if } C_{s\ i}^2 = C_s^2 \text{ for } i = J+1, J+2, \dots, M, \\ 1, & \text{otherwise.} \end{cases}\end{aligned}\tag{3.11}$$

The marginal probability distribution function of the subnetwork σ or σ^* can be obtained by using the maximum entropy analysis of the GE(n)/GE(n)/1/N queue with restricted arrival process (Kouvatsos and Almond, 1988b) and various performance parameters can be calculated. Note that the process of aggregating an arbitrary subnet of service centres with known occupancy in a queueing network leads to a much more efficient computational form for a parametric analysis of the remainder of the network.

Norton-Maximum Entropy Algorithm

This algorithm consists of six steps as follows.

- Step 1: Construct a "shorted" subnetwork 2 (σ^*), a cyclic queueing network (see Figure 8).
- Step 2: For $n = 1, 2, \dots, N$,
 solve aggregate network by applying universal maximum entropy algorithm (Kouvatsos, 1986b) to get the throughput $\chi_1(n)$ and set $\hat{\mu}_1(n)$ equal to $\chi_1(n)$.
- Step 3: Estimate the squared coefficient of variation,

$$\hat{C}_1^2 = \begin{cases} C_s^2, & \text{if } C_{s_i}^2 = C_s^2 \text{ for all } i, i = 1, 2, \dots, M, \\ 1, & \text{otherwise.} \end{cases}$$
- Step 4: Repeat Steps 1-3 for the "shorted" subnetwork 1 (σ) (see Figure 9).
- Step 5: Construct a flow equivalent load dependent two stage cyclic queueing system (see Figure 10),
 set $\hat{\mu}_1(N-n) = \chi_1(N-n)$ and $\hat{\mu}_2(n) = \chi_2(n)$, for $n = 1, 2, \dots, N$,
 set $\hat{C}_1^2$ and $\hat{C}_2^2$ as in [3.11].
- Step 6: Obtain the performance parameters of STAGE 2 by using the maximum entropy analysis of the $GE(n)/GE(n)/1/N$ or $GE(n)/GE/c/N$ queue with restricted arrival process.

Numerical Examples

In this section, numerical results are presented to assess the credibility of the Norton-maximum entropy approach. The results of five examples using Norton-maximum entropy approach are compared with

those obtained by simulation (Russell, 1989) and the universal maximum entropy algorithm (Kouvatsos, 1986b). In the results, N-ME represents the Norton-maximum entropy approach, S represents the simulation done on the original general queueing networks while ME represents the universal maximum entropy algorithm (Kouvatsos, 1986b).

The first three examples consist of single server centres while the last two consist of single server and multiserver centres. The general queueing network for the first example is as depicted in Figure 11 and the results are presented in Table 1 and Figure 12. The second example which consists of six single server centres is as depicted in Figure 13 and the corresponding results are depicted in Figure 14 and Table 2. The results for the third example which again consists of single server centres (see Figure 15) are depicted in Figure 16 and Table 3. As can be seen, the results obtained from the Norton-maximum entropy approach is comparable to those obtained by simulation and universal maximum entropy algorithm.

The general queueing network used in the fourth example is as depicted in Figure 17. This network can similarly be reduced to a flow equivalent two stage cyclic queue as shown in Figure 18. The results for this example are as shown in Table 4 which can be solved by using the maximum entropy analysis of the $GE(n)/GE/c/N$ queue with restricted arrival process (Kouvatsos and Almond, 1988b). In the tables, S_1 represents the simulation done on the original network while S_2 represents the simulation done on the flow equivalent load dependent two stage centres cyclic queue. The last example is as depicted in Figure 19 and this can again be reduced to two stage cyclic queue as

shown in Figure 20 and the results are as shown in Table 5 and Figure 21. From the results of the last two examples, again it is observed that the Norton-maximum entropy approach is comparable in accuracy to that of simulation and the universal maximum entropy algorithm. This is a typical behaviour in many experiments which implies that the marginal probability density function $p_{\sigma^*}(n)$ and various performance parameters of the composite centres C^* of the flow equivalent load dependent two stage GE type cyclic queueing network are relatively close to those in the original GE type network. Note that exact results are obtained in the case of exponential type queueing networks (Chandy *et al.*, 1975a).

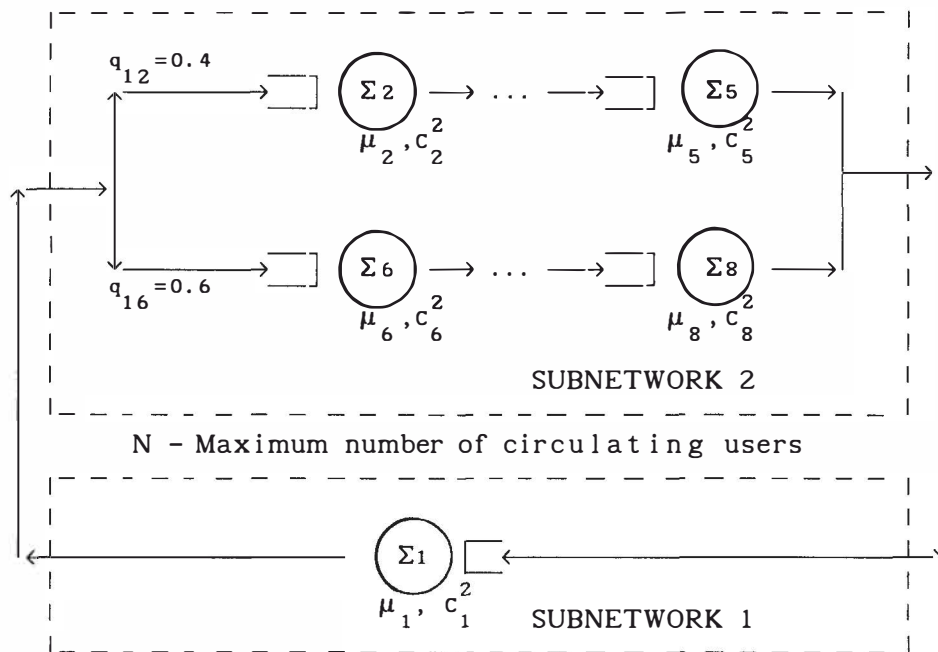


Figure 11 : A Closed Queueing Network with Fixed Routing
for $M = 8$, $\mu_i = 8$, $i = 2, 3, \dots, 8$,
 $C_i^2 = 2$, $i = 1, 2, \dots, 8$

Table 1

Comparison between Simulation (S), Maximum Entropy (ME) and Norton-ME (N-ME) for $M = 8$, $\mu_1 = 8$, $i = 2, 3, \dots, 8$, $C_1^2 = 2$, $i = 1, 2, \dots, 8$

μ_1	N	Time Delay(Subnetwork 2)			Throughput(Subnetwork 2)		
		S	ME	N-ME	S	ME	N-ME
3	2	0.4947	0.5030	0.4950	1.9644	1.9743	1.9283
	4	0.5802	0.5960	0.5834	2.6224	2.6161	2.5597
	6	0.6376	0.6407	0.6326	2.8082	2.8703	2.8260
	8	0.6668	0.6612	0.6585	2.9417	2.9609	2.9345
	10	0.6658	0.6707	0.6710	2.9728	2.9892	2.9764
5	2	0.4994	0.5046	0.5029	2.5380	2.5404	2.4622
	4	0.6124	0.6239	0.6170	3.6208	3.6495	3.5417
	6	0.7067	0.7068	0.6995	4.2529	4.2659	4.1643
	8	0.7604	0.7633	0.7596	4.5561	4.6186	4.5309
	10	0.7913	0.8001	0.8021	4.7405	4.8127	4.7437
7	2	0.5051	0.5065	0.5075	2.8353	2.8635	2.7705
	4	0.6414	0.6400	0.6366	4.2210	4.2640	4.1261
	6	0.7381	0.7479	0.7412	5.1402	5.1432	5.0092
	8	0.8269	0.8368	0.8291	5.6882	5.7393	5.6162
	10	0.9178	0.9095	0.9032	6.1197	6.1541	6.0423

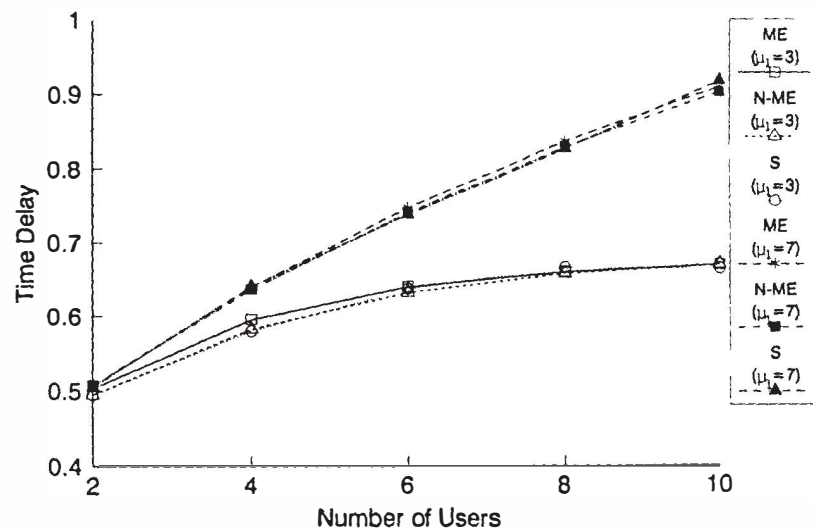


Figure 12 : Time Delay versus Number of Users

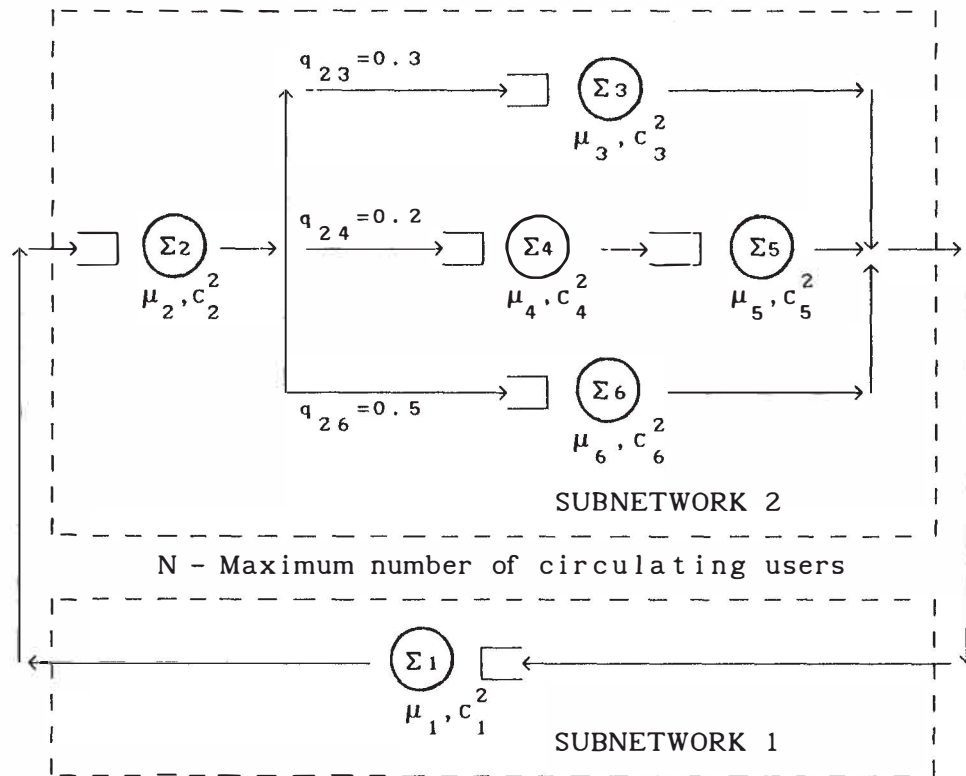


Figure 13 : A Closed Queueing Network for $M = 6$,
 $\mu_2 = 8$, $\mu_3 = \mu_4 = \mu_5 = \mu_6 = 10$,
 $C_1^2 = 1$, $C_2^2 = 0.5$, $C_3^2 = C_4^2 = C_5^2 = C_6^2 = 5$

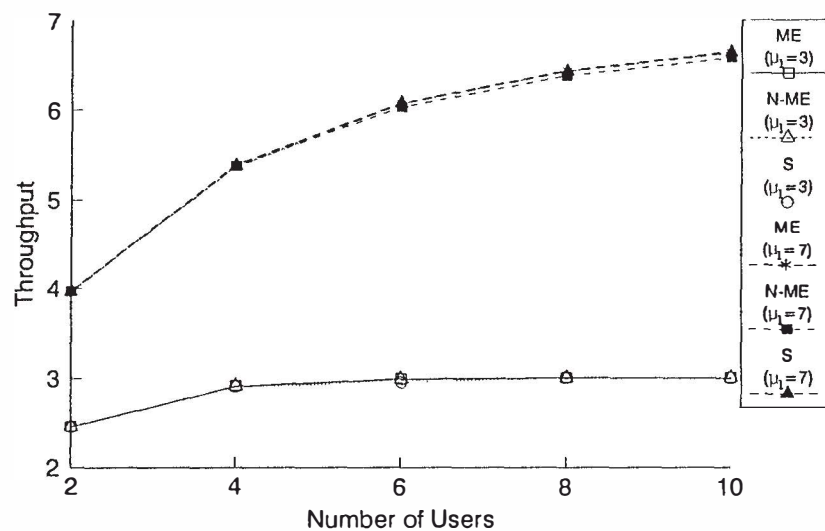


Figure 14 : Throughput versus Number of Users

Table 2

Comparison between Simulation (S), Maximum Entropy (ME) and Norton-ME (N-ME) for $M = 6$, $\mu_2 = 8$, $\mu_3 = \mu_4 = \mu_5 = \mu_6 = 10$, $C_1^2 = 1$, $C_2^2 = 0.5$, $C_3^2 = C_4^2 = C_5^2 = C_6^2 = 5$

μ_1	N	Time Delay(Subnetwork 2)			Throughput(Subnetwork 2)		
		S	ME	N-ME	S	ME	N-ME
3	2	0.2887	0.2855	0.2875	2.4574	2.4607	2.4603
	4	0.3277	0.3274	0.3339	2.9021	2.9105	2.9100
	6	0.3353	0.3401	0.3490	2.9497	2.9866	2.9862
	8	0.3413	0.3429	0.3527	3.0050	2.9981	2.9980
	10	0.3428	0.3433	0.3535	2.9951	2.9998	2.9997
5	2	0.3028	0.2979	0.3002	3.3632	3.3910	3.3895
	4	0.3843	0.3766	0.3875	4.4217	4.4245	4.4089
	6	0.4207	0.4250	0.4465	4.7540	4.7954	4.7738
	8	0.4548	0.4513	0.4824	4.9106	4.9293	4.9121
	10	0.4778	0.4642	0.5023	5.0343	4.9762	4.9657
7	2	0.3078	0.3062	0.3083	3.9648	3.9709	3.9700
	4	0.4183	0.4150	0.4262	5.3794	5.4017	5.3778
	6	0.5109	0.5102	0.5345	6.0688	6.0750	6.0267
	8	0.5981	0.5924	0.6327	6.4202	6.4341	6.3718
	10	0.6585	0.6623	0.7206	6.6324	6.6416	6.5742

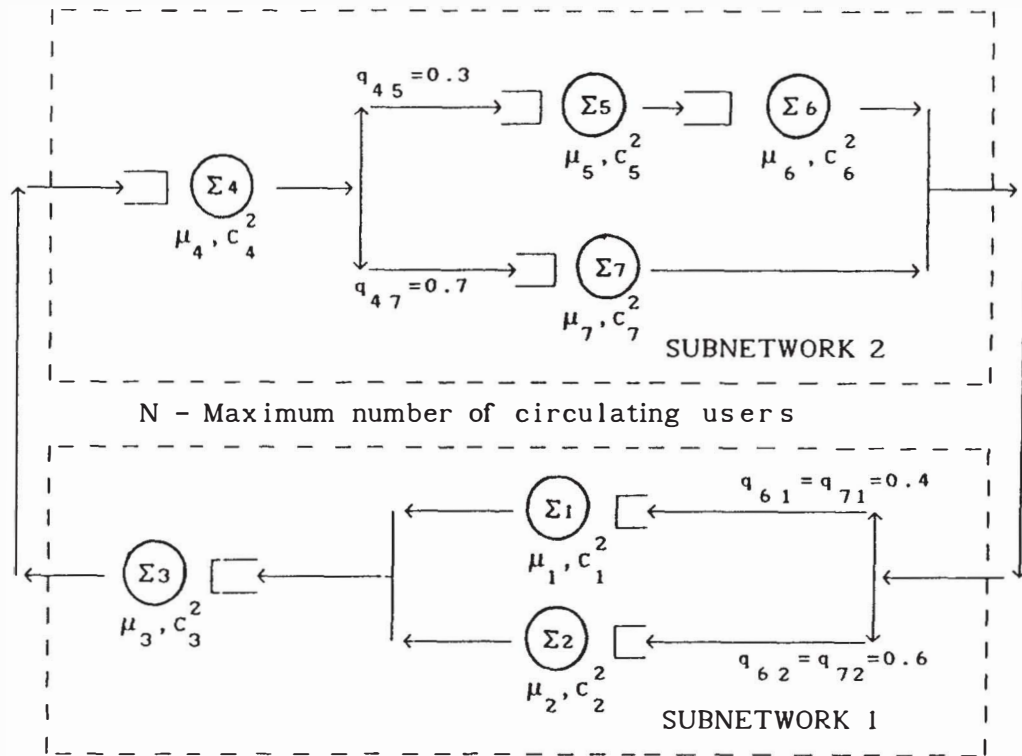


Figure 15 : A Closed Queueing Network for $M = 7$, $\mu_3 = 8$, $\mu_1 = 10$,
 $i = 1, 2, 4, \dots, 7$, $C_1^2 = C_s^2$, $i = 1, 2, \dots, 7$

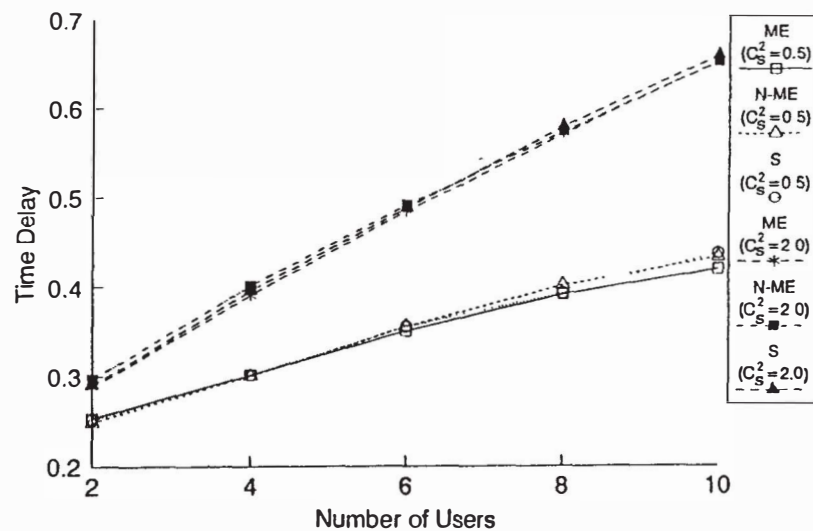


Figure 16: Time Delay versus Number of Users

Table 3

Comparison between Simulation (S), Maximum Entropy (ME)
and Norton-ME (N-ME) for $M = 7$, $\mu_3 = 8$, $\mu_i = 10$,
 $i = 1, 2, 4, \dots, 7$, all $C_i^2 = C_s^2$, $i = 1, 2, \dots, 7$

C_s^2	N	Time Delay(Subnetwork 2)			Throughput(Subnetwork 2)		
		S	ME	N-ME	S	ME	N-ME
0.5	2	0.2507	0.2529	0.2478	3.9483	3.9088	4.0308
	4	0.3016	0.3014	0.2997	6.1787	6.1327	6.3116
	6	0.3550	0.3502	0.3551	7.2389	7.2014	7.3150
	8	0.3919	0.3904	0.4004	7.6831	7.6677	7.7256
	10	0.4361	0.4183	0.4315	7.9237	7.8642	7.8902
1.0	2	0.2675	0.2667	0.2667	3.7292	3.7211	3.7211
	4	0.3409	0.3402	0.3402	5.6143	5.5915	5.5915
	6	0.4139	0.4108	0.4108	6.6439	6.6006	6.6006
	8	0.4793	0.4756	0.4756	7.1771	7.1681	7.1681
	10	0.5388	0.5323	0.5323	7.4814	7.4966	7.4966
2.0	2	0.2902	0.2877	0.2971	3.4138	3.4672	3.3235
	4	0.3954	0.3905	0.4010	4.9167	4.9732	4.7735
	6	0.4872	0.4836	0.4906	5.7973	5.8686	5.6859
	8	0.5782	0.5697	0.5727	6.3606	6.4552	6.3022
	10	0.6561	0.6498	0.6489	6.7430	6.8588	6.7343
5.0	2	0.3262	0.3268	0.3491	3.0136	3.0600	2.8303
	4	0.4926	0.4771	0.5140	4.0169	4.1354	3.7569
	6	0.6108	0.6052	0.6431	4.7145	4.8386	4.4350
	8	0.7482	0.7218	0.7561	5.2301	5.3538	4.9625
	10	0.8572	0.8306	0.8595	5.5646	5.7526	5.3866

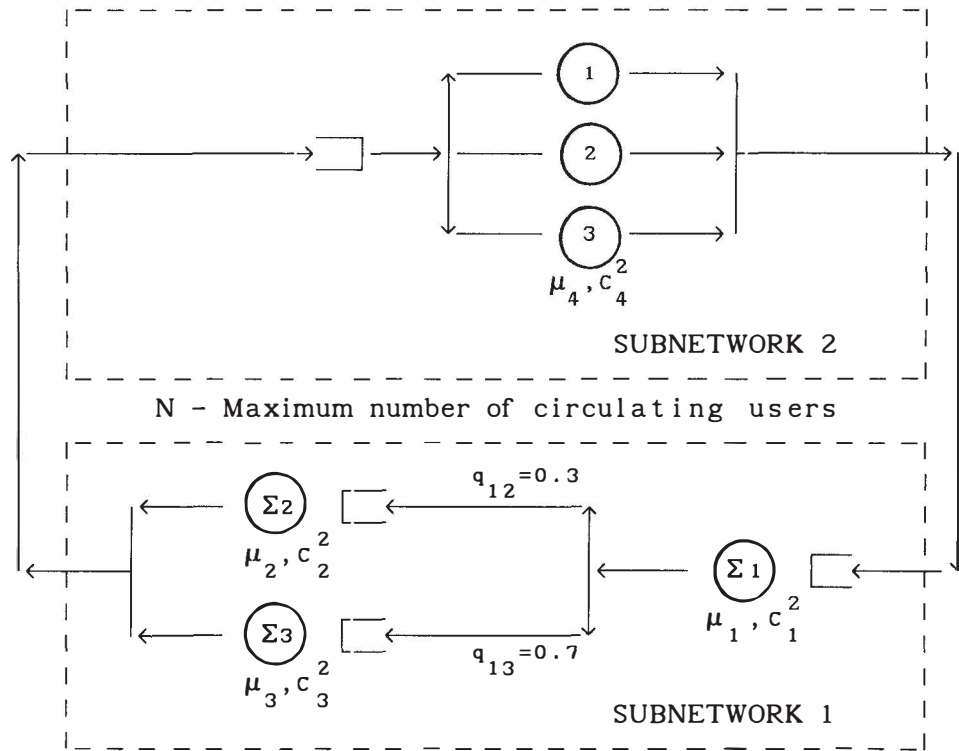


Figure 17 : A Closed Queueing Network with Fixed Routing for $M = 4$,
 $c = 3$, $\mu_1 = 7$, $\mu_2 = \mu_3 = 10$, $C_i^2 = C_s^2$, $i = 1, 2, \dots, 4$

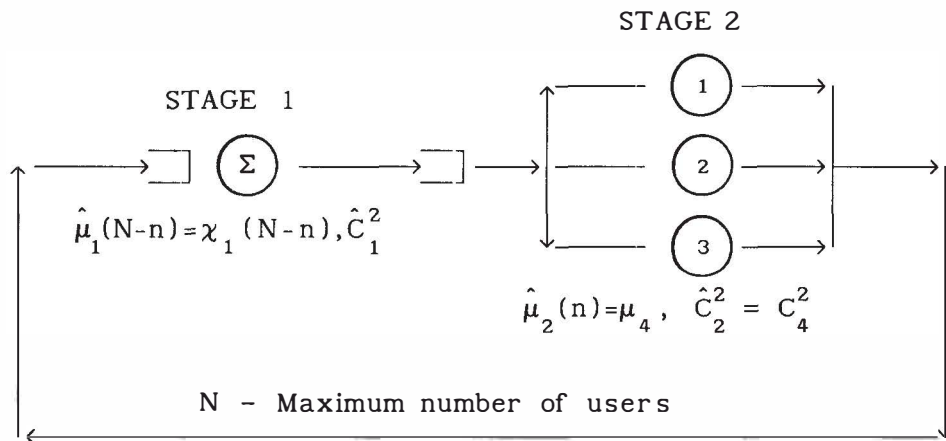


Figure 18 : A Flow Equivalent Load Dependent Two Stage
 Multiserver Cyclic Queueing System

Table 4

Comparison between Simulation (S_1, S_2) and Norton-ME (N-ME)
 for $M = 4, \mu_1 = 7, \mu_2 = \mu_3 = 10, C_i^2 = C_s^2, i = 1, 2, 3, 4$

C_s^2	μ_4	N	Time Delay(Subnetwork 2)			Throughput(Subnetwork 2)		
			S_1	S_2	N-ME	S_1	S_2	N-ME
0.5	4	3	0.2496	0.2494	0.2500	5.1412	4.4431	5.1960
		6	0.2656	0.2625	0.2607	6.7841	6.6168	6.8461
		9	0.2758	0.2744	0.2669	6.9538	6.9912	6.9923
		12	0.2753	0.2743	0.2675	6.9585	7.0082	6.9997
	8	3	0.1241	0.1247	0.1250	6.0324	5.1300	6.0925
		6	0.1251	0.1252	0.1249	6.9678	6.9211	6.9722
		9	0.1254	0.1253	0.1249	7.0090	7.0039	6.9994
		12	0.1254	0.1253	0.1249	7.0091	7.0060	7.0000
5.0	4	3	0.2465	0.2560	0.2500	4.3177	3.9179	4.1874
		6	0.3211	0.3253	0.3270	5.2596	4.7231	5.1454
		9	0.3479	0.3910	0.3941	5.8120	5.3275	5.7202
		12	0.4271	0.4397	0.4538	6.2551	5.8081	6.2301
	8	3	0.1210	0.1280	0.1250	4.7642	4.3468	4.7104
		6	0.1431	0.1483	0.1491	5.7730	5.2013	5.6716
		9	0.1438	0.1668	0.1649	6.1582	5.8329	6.2010
		12	0.1559	0.1743	0.1760	6.7083	6.3470	6.7382

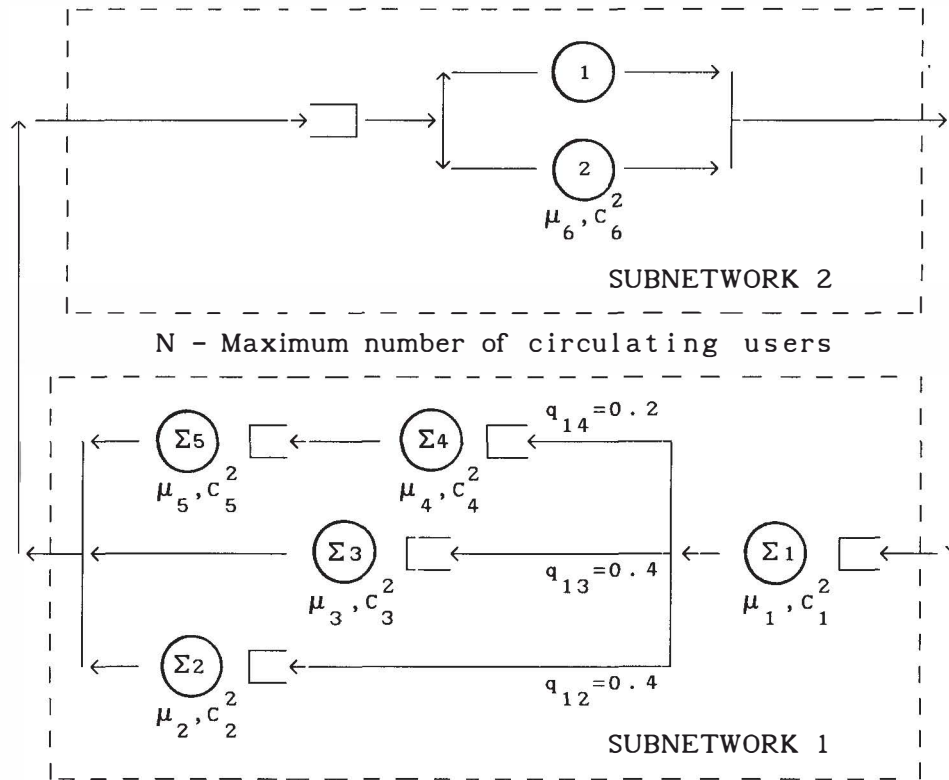


Figure 19 : A Closed Queueing Network with Fixed Routing for
 $M = 6, c = 2, \mu_1 = 8, \mu_i = 10, i = 2, 3, \dots, 5,$
 $C_1^2 = C_s^2, i = 1, 2, \dots, 6$

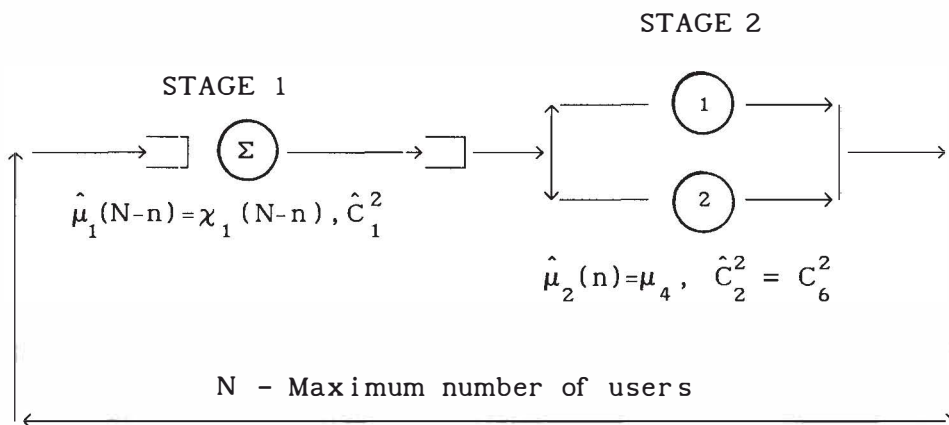


Figure 20 : A Flow Equivalent Load Dependent Two Stage
 Multiserver Cyclic Queueing System

Table 5

Comparison between Simulation (S_1 , S_2) and Norton-ME (N-ME)
 for $M = 6$, $\mu_1 = 8$, $\mu_i = 10$, $i = 2, 3, \dots, 5$, $C_i^2 = C_s^2$,
 $i = 1, 2, \dots, 6$

C_s^2	μ_6	N	Time Delay(Subnetwork 2)			Throughput(Subnetwork 2)		
			S_1	S_2	N-ME	S_1	S_2	N-ME
0.5	4	2	0.2489	0.2494	0.2500	3.8311	3.3416	4.3832
		5	0.3534	0.3261	0.3416	6.6235	6.1271	6.6859
		8	0.5265	0.4706	0.5032	7.3020	7.1544	7.3297
		11	0.7109	0.6391	0.6908	7.5688	7.5132	7.5539
		14	0.8682	0.7847	0.8752	7.7284	7.6195	7.6651
	6	2	0.1665	0.1663	0.1667	4.4847	3.6906	5.2806
		5	0.2002	0.1875	0.1880	7.4691	6.9986	7.5248
		8	0.2265	0.2072	0.2081	8.0286	7.8781	7.9565
		11	0.2348	0.2143	0.2128	8.1028	7.9795	7.9965
		14	0.2328	0.2158	0.2134	8.1016	7.9928	7.9951
2.0	4	2	0.2466	0.2504	0.2500	3.6534	3.2630	3.6140
		5	0.3890	0.3874	0.4037	5.5493	4.9471	5.4986
		8	0.5675	0.5274	0.5739	6.2855	5.9581	6.2902
		11	0.7168	0.6886	0.7521	6.8894	6.4701	6.7057
		14	0.8964	0.8554	0.9344	7.0168	6.8148	6.9591
	6	2	0.1668	0.1669	0.1667	4.1745	3.6389	4.1668
		5	0.2237	0.2313	0.2404	6.4597	5.6290	6.3441
		8	0.2687	0.2917	0.3050	7.2128	6.8428	7.2127
		11	0.3438	0.3459	0.3574	7.5830	7.4342	7.6036
		14	0.3867	0.3966	0.3970	7.7103	7.7342	7.7933

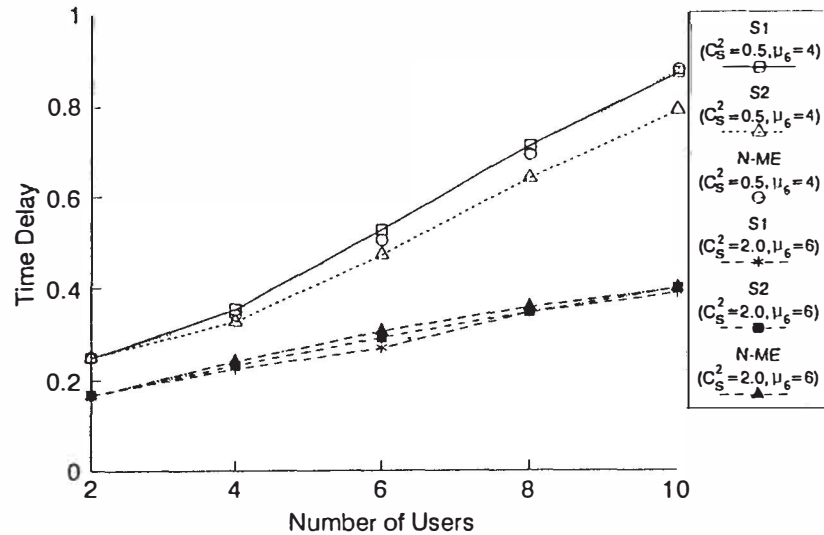


Figure 21 : Time Delay versus Number of Users

Summary

An extension to Norton's theorem for queueing networks using the universal maximum entropy algorithm has been proposed for the analysis of large general closed queueing networks. This new approach can be used to reduce a complex system to a two stage general composite centres with load dependent service rates and has an advantage over the direct application of the universal maximum entropy whereby the study of a subset of queueing centres of interest can be done without repeatedly solving the entire network. Therefore substantial computational effort can be reduced. From the results obtained, it is observed that the new approach is comparable in accuracy to that of simulation and universal maximum entropy approach.

CHAPTER IV

SINGLE SERVER GENERAL QUEUEING NETWORKS WITH MULTICLASS USERS

Introduction

In this chapter, a new approximate analytical solution for the joint queue length distribution of the single server general closed queueing network with multiclass users is proposed. The single server general closed queueing network with multiclass users can be used to model several practical situations such as the machine repairman problem (Allen, 1978), multiprocessor computer systems (Marson and Gregoretti, 1981) and the central server system (Sauer and Chandy, 1975; Foxley and Salman, 1981) where it is assumed that a fixed number of users (i.e. programs) traverse a closed queueing network consisting of the central processing unit (CPU) and several input/output (I/O) devices. The single server general closed queueing network with multiclass users can also be used to model the decentralised optimal flow control of computer communication networks with single transmission link and multiple controllers (Hsiao and Lazar, 1984; 1985).

This chapter is organised as follows. A detailed description of the single server general closed queueing network with multiclass users is presented followed by the derivation of the analytic maximum entropy formulation of the single server general closed queueing network with multiclass users. This derivation is based on the known heuristic arguments and earlier results regarding the study of the stable $G/G/1$ and $G/G/1/N$ queues. The one-step maximum entropy

recursions solution is then presented followed by the solution to load dependent servers. Numerical examples will be provided to test the credibility of the new solution followed by a summary of this chapter.

Description of the Network

The general queueing network we now consider is as shown in Figure 22. This network which is a generalisation of the queueing network used by Hsiao and Lazar (1984; 1985; 1990) in their studies on flow control, consists of $K+1$ infinite queueing centres and K classes of users. Centre 0 is modelled as a general server with fixed service rate μ , constant coefficient of variation C_μ^2 and a FCFS service discipline. Users belonging to either class join the common queue at centre 0 to wait for service. After receiving service at centre 0, users of class i ($i = 1, 2, \dots, K$) proceed to centre i for their next

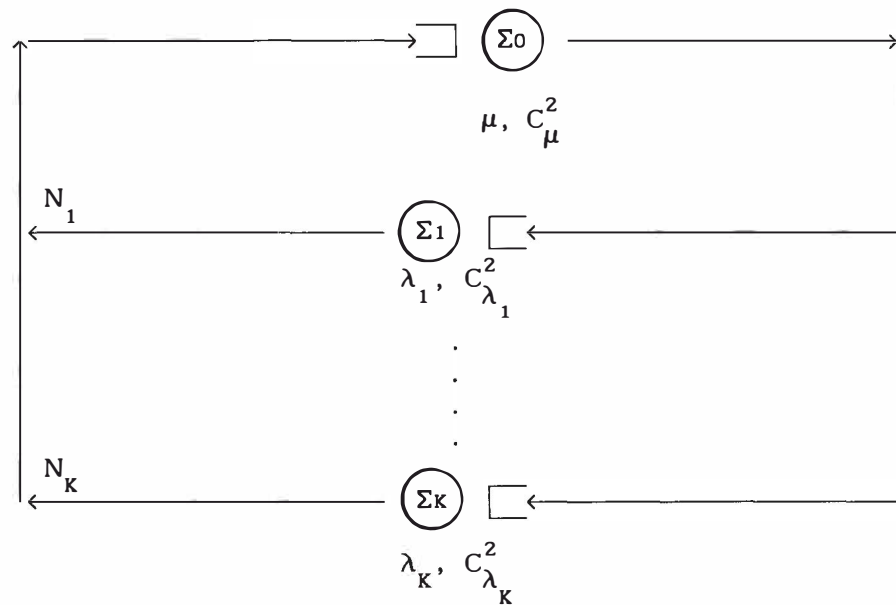


Figure 22 : Queueing Model for Multiclass Users and Single Server Centres

service. Centre i accept users of class i at a rate λ_i users per second and constant coefficient of variation of C_{λ_i} on a FCFS basis. The maximum number of users is N , $N = N_1 + N_2 + \dots + N_K$, where N_i is the maximum number of class i users.

Maximum Entropy Formulations

In this section, the maximum entropy analytical approach used in solving a mixed class (i.e. K closed) general queueing network is presented. The model depicted in Figure 22 is a K closed class general queueing network with a product form solution where n_i being the number of class i ($i = 1, 2, \dots, K$) users in the queue of centre 0 respectively. λ_i is the corresponding arrival rates. Note that since the routing depends on the class of users, the general closed network model is an extended Baskett *et al.* (1975) network and can be analysed by using the maximum entropy approximation approach. Let $\underline{n} = (n_1, n_2, \dots, n_K)$ be the state of centre 0 where $n_i, n_i = 0, 1, \dots, N_i$, and N_i is the total number of users of class i . Furthermore let $N = \sum_{i=1}^K N_i$ be the total number of users in the queueing system. The following analysis assumes that the parameters $\lambda_i, \mu, C_{\lambda_i}^2$ and C_{μ}^2 , form a basic set of a *priori* knowledge. Suppose all that is known about the steady state joint probability $p_N(\underline{n})$, $\underline{n} = \{n_1, n_2, \dots, n_K\}$, are the following mean value constraints:

- i. the normalisation condition,

$$\sum_{n_1=0}^{N_1} \dots \sum_{n_K=0}^{N_K} p_N(\underline{n}) = 1 \quad [4.1]$$

- ii. the utilisation, ρ , $0 < \rho < 1$,

written as

$$\sum_{n_1=0}^{N_1} \dots \sum_{n_K=0}^{N_K} h(\underline{n}) p_N(\underline{n}) = \rho, \quad [4.2]$$

$$h(\underline{n}) = \begin{cases} 0, & \text{if } n_1 + n_2 + \dots + n_K = 0, \\ 1, & \text{otherwise,} \end{cases}$$

$$\rho = \sum_{i=1}^K \rho_i, \quad \rho_i = \lambda_i / \mu.$$

- iii. the mean queue length per class $\langle n_i \rangle_{N_i}$, $\rho_i \leq \langle n_i \rangle_{N_i} < N_i$,

$$\sum_{n_1=0}^{N_1} \dots \sum_{n_K=0}^{N_K} n_i p_N(\underline{n}) = \langle n_i \rangle_{N_i}, \quad i = 1, 2, \dots, K, \quad [4.3]$$

$$\text{where } \langle N \rangle_N = \sum_{i=1}^K \langle n_i \rangle_{N_i}.$$

- iv. the flow balance,

$$\left(\sum_{i=1}^K \lambda_i \right) \sum_{n_1=0}^{N_1-1} \sum_{n_2=0}^{\infty} \dots \sum_{n_K=0}^{\infty} p_N(\underline{n}) +$$

$$\left(\sum_{i=2}^K \lambda_i \right) \sum_{n_2=0}^{\infty} \dots \sum_{n_K=0}^{\infty} p_N(N_1, n_2, \dots, n_K) = \mu[1-p(\underline{0})], \quad [4.4]$$

for $n_1 = N_1$, $n_i < N_i$ ($i = 2, 3, \dots, K$),

⋮

$$\begin{aligned}
 & \left(\sum_{i=1}^K \lambda_i \right) \sum_{n_1=0}^{N_1-1} \dots \sum_{n_K=0}^{N_K-1} p_N(\underline{n}) + \lambda_1 \sum_{n_1=0}^{N_1-1} p_N(n_1, N_2, \dots, N_K) + \dots + \\
 & \lambda_K \sum_{n_K=0}^{N_K-1} p_N(N_1, N_2, \dots, n_K) = \mu[1-p(0)], \tag{4.5}
 \end{aligned}$$

for all $n_i = N_i$ ($i = 1, 2, \dots, K$).

Note that there will be a total of $\sum_{k=0}^{K-1} 2^k$ flow balance constraints. This corresponds to a total of $\sum_{k=0}^{K-1} 2^k$ ways the state of centre 0 (i.e. $\underline{n}$) contains at least one $n_i = N_i$. The above statistics may be known numerically via system measurements during finite observations periods or can be determined via known analytic formulae (if available) which are usually expressed in terms of moments of the interarrival and service time distributions. Since the number of constraints is less than the number of feasible states, there is an infinite number of distributions $\{p_N(\underline{n})\}$ subject to these constraints. The problem is which one to choose. The principle of maximum entropy states that of all the distributions satisfying the constraints supplied by the given information, the form of the minimally biased distribution which should be chosen is the one that maximises the system's entropy function

$$H(p) = - \sum_{n_1=0}^{N_1} \dots \sum_{n_K=0}^{N_K} p_N(\underline{n}) \log \{ p_N(\underline{n}) \}. \quad [4.6]$$

The formulation of the approximate maximum entropy queue length distribution of the G/G/1/N queue with K ($K \geq 2$) classes of users will begin from the G/G/1/N queue with single class users as follows. From the results obtained by Kouvatso (1986a; 1988a), the maximum entropy queue length distribution of the G/G/1/N queue with single class users ($K=1$), subject to constraints [4.1]–[4.5] is given by

$$p_N(n_1) = p_N(0) g^{h(n_1)} x^{n_1} y^{f(n_1)}, \quad [4.7]$$

$$f(n_1) = \begin{cases} 0, & \text{if } n_1 < N_1, \\ 1, & \text{if } n_1 = N_1, \end{cases}$$

where $p_N(0)$ is obtained from normalisation and is given by

$$p_N(0) = \frac{1 - x}{1 - x + gx(1 - x^{N-1}) + (1 - x)gyx^N}, \quad [4.8]$$

$$g = \frac{(1 - x) \rho}{(1 - \rho) x}, \quad [4.9]$$

$$x = \frac{\langle N \rangle - \rho}{\langle N \rangle}, \quad [4.10]$$

$$y = \frac{(1 - \rho)}{(1 - x)}, \quad [4.11]$$

where $\langle N \rangle$ is the mean queue length constraint of the single class

$G/G/1$ queue. The one-step transition weight diagram of the $G/G/1/N$ queue with single class users is as depicted in Figure 23. Let us now

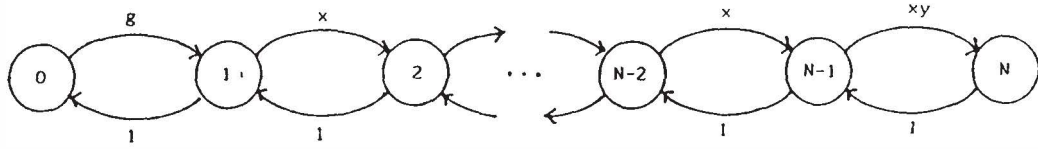


Figure 23 : Transition Weight Diagram of a $G/G/1/N$ Queue with Single Class Users

consider the $G/G/1/N$ queue with two classes (i.e. $K=2$) of users and the one-step transition weight diagram is as depicted in Figure 24.

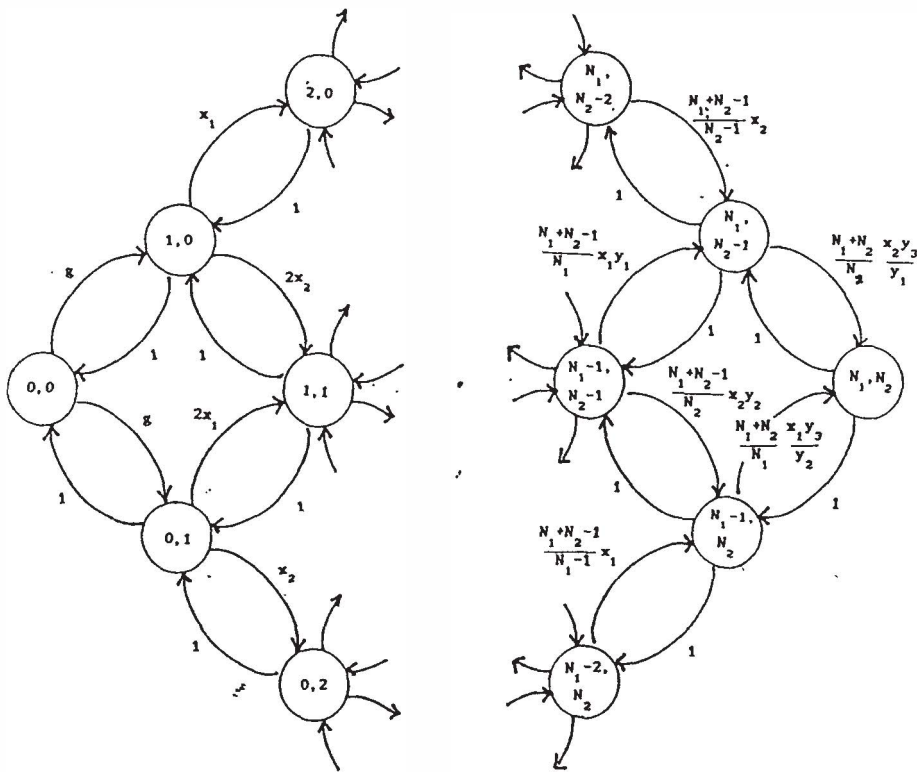


Figure 24 : Transition Weight Diagram of a $G/G/1/N$ Queue with Two Classes of Users

Theorem 4.1

The joint queue length maximum entropy solution $p_{\underline{n}}(\underline{n})$ of a stable G/G/1/N queue with two classes of users under FCFS service discipline that maximises the system's entropy function [4.6] subject to the constraints [4.1]-[4.5] (Othman *et al.*, 1992) is given by

$$p_{\underline{n}}(\underline{n}) = \frac{1}{Z(p)} \binom{n_1+n_2}{n_1} g^{\underline{h}(\underline{n})} x_1^{n_1} x_2^{n_2} y_1^{f_1(\underline{n})} y_2^{f_2(\underline{n})} y_3^{f_3(\underline{n})}, \quad [4.12]$$

$$f_1(\underline{n}) = \begin{cases} 1, & \text{if } n_1 = N_1 \text{ \& } n_2 < N_2, \\ 0, & \text{otherwise,} \end{cases}$$

$$f_2(\underline{n}) = \begin{cases} 1, & \text{if } n_1 < N_1 \text{ \& } n_2 = N_2, \\ 0, & \text{otherwise,} \end{cases}$$

$$f_3(\underline{n}) = \begin{cases} 1, & \text{if } n_1 = N_1 \text{ \& } n_2 = N_2, \\ 0, & \text{otherwise,} \end{cases}$$

where $Z(p)$ is the normalisation constant given by

$$Z(p) = \exp(\beta_0) =$$

$$\sum_{n_1=0}^{N_1} \sum_{n_2=0}^{N_2} \binom{n_1+n_2}{n_1} g^{\underline{h}(\underline{n})} x_1^{n_1} x_2^{n_2} y_1^{f_1(\underline{n})} y_2^{f_2(\underline{n})} y_3^{f_3(\underline{n})}, \quad [4.13]$$

$g = \exp(-\beta_1)$, $x_1 = \exp(-\beta_{21})$, $x_2 = \exp(-\beta_{22})$, $y_1 = \exp(-\beta_{31})$, $y_2 = \exp(-\beta_{32})$, $y_3 = \exp(-\beta_{33})$ and β_1 , β_{21} , β_{22} , β_{31} , β_{32} , β_{33} are the Langrangian multipliers corresponding to constraints [4.2]-[4.5], respectively.

Proof

The maximisation of the entropy function [4.6] subject to constraints [4.1]–[4.5] can be carried out using the Langrange's method of undetermined multipliers.

By using [4.1] and [4.2], $p_N(\underline{0})$ can be expressed by

$$p_N(\underline{0}) = \frac{1}{Z(p)} = \left[\sum_{n_1=0}^{N_1} \sum_{n_2=0}^{N_2} \binom{n_1+n_2}{n_1} g^{\frac{h(n)}{x_1}} x_1^{n_1} x_2^{n_2} y_1^{\frac{f_1(n)}{y_1}} y_2^{\frac{f_2(n)}{y_2}} y_3^{\frac{f_3(n)}{y_3}} \right]^{-1} \quad [4.14]$$

Corollary

Assuming that $x_1, x_2 < 1$, x_1 is invariant of the buffer size N_1 and x_2 is invariant to the buffer size N_2 and allowing $N_1, N_2 \rightarrow \infty$, [4.12] and [4.14] reduce to

$$p_N(\underline{n}) = p(\underline{0}) \binom{n_1+n_2}{n_1} g^{\frac{h(n)}{x_1}} x_1^{n_1} x_2^{n_2}, \quad [4.15]$$

$$p(\underline{0}) = 1 + \frac{gx_2}{(1-x_2)} + \frac{gx_1}{(1-x_2)(1-x_1-x_2)} = 1 - \rho, \quad [4.16]$$

where $\rho = (\lambda_1 + \lambda_2)/\mu$, ($\rho < 1$). These expressions represents the maximum entropy solution of the corresponding infinite capacity G/G/1 queue at equilibrium with two classes of users subject to utilisation ρ and

mean queue lengths $\langle n_1 \rangle$ and $\langle n_2 \rangle$ constraints with Lagrangian coefficients g and x_i given by

$$g = \frac{(1 - x) \rho}{(1 - \rho) x}, \quad [4.17]$$

$$x_i = \frac{\langle n_i \rangle (1 - x) x}{\rho}, \quad i = 1, 2, \quad [4.18]$$

$$x = \frac{\sum_{i=1}^2 \langle n_i \rangle - \rho}{\sum_{i=1}^2 \langle n_i \rangle} = x_1 + x_2, \quad x_1 = \pi x, \quad \pi = \lambda_1 / (\lambda_1 + \lambda_2), \quad [4.19]$$

Centre 0 can be seen as an infinite G/G/1 general queueing system with two classes of users if the number of outstanding class 1 and 2 users is less than N_1 and N_2 respectively. Centre 0 can also be seen as being a restricted G/G/1/ N_1 finite capacity general queueing system having two classes of users when the number of outstanding class 2 users is always less than N_2 . The marginal probability, $p_{N_1}(n_1, \cdot)$, $n_1 = 0, 1, \dots, N_1$, $n_2 = 0, 1, \dots$, is given by

$$\begin{aligned} p_{N_1}(n_1, \cdot) &= \sum_{n_2=0}^{\infty} p_N(n_1, n_2) \\ &= p_N(0) \sum_{n_2=0}^{\infty} \binom{n_1+n_2}{n_1} g^{\frac{h(n)}{x_1} \frac{n_1}{x_2} \frac{n_2}{y_1} \frac{f_1(n)}{y_1}} \end{aligned} \quad [4.20]$$

After some manipulation, $p_{N_1}(n_1, \cdot)$ can be written as

$$p_{N_1}(n_1, \cdot) = p_N(0) G x_1^{\frac{n_1}{x_1} \frac{f_1(n)}{y_1}}, \quad [4.21]$$

where

$$G = \begin{cases} 1 + \frac{gx_2}{(1-x_2)}, & \text{if } n_1 = 0, \\ \frac{g}{(1-x_2)^{n_1+1}}, & \text{if } n_1 > 0. \end{cases}$$

y_1 can easily be computed from [4.21] and the flow balance constraint and is given by

$$y_1 = \frac{(1-\rho)(1-x_2)}{(1-x)(1-\rho_2)}, \quad [4.22]$$

where $\rho = \rho_1 + \rho_2$, $\rho_1 = \lambda_1/\mu$ and $\rho_2 = \lambda_2/\mu$. Centre 0 can also be seen as a restricted $G/G/1/N_2$ finite capacity queueing system having two classes of users when the number of outstanding class 1 users is always less than N_1 . Due to the symmetric behaviour of class 1 and 2 users, y_2 can be expressed as

$$y_2 = \frac{(1-\rho)(1-x_1)}{(1-x)(1-\rho_1)}, \quad [4.23]$$

Finally, if the total number of outstanding class 1 and 2 users is N , centre 0 can be seen as a restricted $G/G/1/N$ finite capacity queueing system having two classes of users whereby centre 1 and 2 model as the input streams. The flow balance constraint and [4.14] in this case can be written as

$$(\lambda_1 + \lambda_2)[1 - p_N(N)] = \mu[1 - p_N(0)], \quad [4.24]$$

$$p_N(0) = \frac{1 - x}{1 - x + gx(1 - x^{N-1}) + (1 - x)gy_3x^N}, \quad [4.25]$$

It can be verified by using [4.24] and [4.25], the Langrangian coefficient y_3 is given by

$$y_3 = \frac{(1 - \rho)}{(1 - x)}. \quad [4.26]$$

In general, the above solutions can be extended to the G/G/1/N queue with K (>2) classes of users by assuming that the Langrange coefficients g and x have the same form as in the G/G/1/N queue with single class and two classes of users .

Theorem 4.2

The joint maximum entropy solution $p_N(\underline{n})$ of a stable G/G/1/N queue with K (≥ 2) classes of users under FCFS service discipline that maximises the system's entropy function [4.6] subject to the constraints [4.1]-[4.5] is given by

$$p_N(\underline{n}) = \frac{1}{Z(p)} \frac{\left[\sum_{i=1}^K n_i \right]!}{\left[\prod_{i=1}^K n_i! \right]} g^{\underline{h}(\underline{n})} \left[\prod_{i=1}^K x_i^{n_i} \right] y_1^{f(1)} \dots y_t^{f(t)}, \quad [4.27]$$

$$n_i = 0, 1, \dots, N_i$$

where $Z(p)$ is the normalisation constant given by

$$Z(p) = \exp(\beta_0) =$$

$$\sum_{n_1=0}^{N_1} \dots \sum_{n_K=0}^{N_K} \frac{\left[\sum_{i=1}^K n_i \right]!}{\left[\prod_{i=1}^K n_i! \right]} g^{\sum_{i=1}^K n_i} y_1^{f(1)} \dots y_t^{f(t)} \quad [4.28]$$

$$f(j) = \begin{cases} 1, & \text{if } j \text{ corresponds to one of the ways that centre } 0 \\ & \text{contains at least one } n_i = N_i, i = 1, 2, \dots, K, \\ 0, & \text{otherwise,} \end{cases}$$

$$j = 1, 2, \dots, t = \sum_{k=0}^{K-1} 2^k,$$

$$g = \exp(-\beta_1) = \frac{(1-x)\rho}{(1-\rho)x}, \quad [4.29]$$

$$x_i = \exp(-\beta_{2i}) = \frac{\langle n_i \rangle (1-x)x}{\rho}, \quad i = 1, 2, \dots, K, \quad [4.30]$$

$$x = \frac{\sum_{i=1}^K \langle n_i \rangle - \rho}{\sum_{i=1}^K \langle n_i \rangle} = \frac{\sum_{i=1}^K x_i}{\sum_{i=1}^K x_i}, \quad x_i = \pi x, \quad \pi = \lambda_i / (\lambda_1 + \dots + \lambda_K),$$

$$y_j = \exp(-\beta_{3j}) = \frac{(1-\rho) \left[1 - \sum_{\substack{k=1 \\ k \neq \text{finite capacity centre}}}^K x_k \right]}{(1-x) \left[1 - \sum_{\substack{k=1 \\ k \neq \text{finite capacity centre}}}^K \rho_k \right]}, \quad [4.31]$$

where β_0 , β_1 , β_{2i} and β_{3j} are the Lagrangian multipliers corresponding to constraints [4.1] - [4.5] respectively.

Proof

The maximisation of the entropy function [4.6], subject to constraints [4.1] - [4.5] can be carried out by using the Lagrange's method of undetermined multipliers leading to the solution [4.27] - [4.31].

By using [4.1] and [4.2], $p_N(\underline{0})$ can be expressed by

$$p_N(\underline{0}) = \frac{1}{Z(p)} = \left[\sum_{n_1=0}^{N_1} \dots \sum_{n_K=0}^{N_K} \frac{\left[\sum_{i=1}^K n_i \right]!}{\left[\prod_{i=1}^K n_i! \right]} g^{\underline{h(n)}} \left[\prod_{i=1}^K x_i^{n_i} y_1^{f(1)} \dots y_t^{f(t)} \right]^{-1} \quad [4.32]$$

Corollary

Assuming that $x_i < 1$, x_i is invariant of the buffer size N_i and allowing $N_i \rightarrow \infty$ ($i = 1, 2, \dots, K$), [4.27] and [4.32] reduce to

$$p(\underline{n}) = p(\underline{0}) \frac{\left[\sum_{i=1}^K n_i \right]!}{\left[\prod_{i=1}^K n_i! \right]} g^{\underline{h(n)}} \left[\prod_{i=1}^K x_i^{n_i} \right], \quad [4.33]$$

$$p(\underline{0}) = \left[1 + \frac{gx_K}{(1-x_K)} + \frac{gx_{K-1}}{(1-x_K)(1-x_{K-1}-x_K)} + \dots + \frac{gx_1}{(1-x_2-\dots-x_K)(1-x_1-x_2-\dots-x_K)} \right]^{-1} = 1 - \rho, \quad [4.34]$$

where $\rho = (\lambda_1 + \dots + \lambda_K)/\mu$, ($\rho < 1$). These expressions represents the maximum entropy solution of the corresponding infinite capacity G/G/1 queue at equilibrium with K classes of users subject to utilisation ρ and mean queue lengths $\langle n_i \rangle$ constraints with Lagrangian coefficients g and x_i given by [4.29] and [4.30]. The versatile form of the joint maximum entropy solution [4.27] will be used in the next section to establish the one-step maximum entropy recursions.

One-Step Maximum Entropy Recursions

From the joint maximum entropy solution [4.27], the following one-step maximum entropy recursions are established.

$$p_N(1, 0, \dots, 0) = p_N(0) g x_1, \quad [4.35]$$

$$\vdots$$

$$p_N(0, 0, \dots, 1) = p_N(0) g x_K, \quad [4.36]$$

$$\vdots$$

$$p_N(n_1, 0, \dots, 0) = x_1 y_1^{f(1)} p_N(n_1 - 1, 0, \dots, 0), \quad [4.37]$$

$$n_1 = 2, 3, \dots, N_1,$$

$$\vdots$$

$$p_N(0, \dots, 0, n_K) = x_K y_t^{f(t)} p_N(0, \dots, 0, n_K - 1), \quad n_K = 2, 3, \dots, N_K, \quad [4.38]$$

$$p_N(\underline{n}) = \begin{cases} \frac{(n_1 + \dots + n_K)}{n_1} x_1 Z(y) p(n_1 - 1, n_2, \dots, n_K), \\ \vdots \\ \frac{(n_1 + \dots + n_K)}{n_K} x_K Z(y) p(n_1, \dots, n_{K-1}, n_K - 1), \end{cases} \quad [4.39]$$

where $Z(y)$ represents the various combination of the Langrangian coefficients, y , that will occur when $n_i = N_i$. Recursions [4.35]-[4.39] enable analyst to study more general operational or stochastic K resource queues with finite buffer, N , for which operational Birth-Death transitions (Buzen and Denning, 1980) or classical queueing theory (Kleinrock, 1975a; Allen, 1978) no longer effectively apply. Due to the complexity in writing down the one-step maximum entropy recursion solutions when using either the bulk or the limiting interpretation of the GE distribution (Kouvatsos, 1988a), let us now for simplicity, consider for the case $K = 2$. [4.35]-[4.39] can now be written in the form given by (Ku-Mahamud and Othman, 1992)

$$p_N(1, 0) = g x_1 p_N(0, 0), \quad [4.40]$$

$$p_N(0, 1) = g x_2 p_N(0, 0), \quad [4.41]$$

$$p_N(n_1, 0) = x_1 y_1^{f_1(n)} p_N(n_1 - 1, 0), \quad n_1 = 2, 3, \dots, N_1, \quad [4.42]$$

$$p_N(0, n_2) = x_2 y_2^{\frac{f_2(n)}{2}} p_N(0, n_2-1), \quad n_2 = 2, 3, \dots, N_2, \quad [4.43]$$

$$p_N(n_1, n_2) = \begin{cases} \frac{\binom{n_1+n_2}{n_1}}{x_1 y_1^{\frac{f_1(n)}{2}}} (y_3/y_2)^{\frac{f_3(n)}{3}} p_N(n_1-1, n_2), \\ \frac{\binom{n_1+n_2}{n_2}}{x_2 y_2^{\frac{f_2(n)}{2}}} (y_3/y_1)^{\frac{f_3(n)}{3}} p_N(n_1, n_2-1), \end{cases}$$

$$\begin{aligned} n_1 &= 1, 2, \dots, N_1, \\ n_2 &= 1, 2, \dots, N_2. \end{aligned} \quad [4.44]$$

It can be shown that by using either the bulk or the limiting interpretation of the GE distribution, [4.40]-[4.44] can be written in the form given by

$$p_N(n_1, n_2) = \begin{cases} \frac{\binom{n_1+n_2}{n_1}}{a(n_1-1, n_2)} \frac{b(n_1, n_2)}{b(n_1, n_2)} p_N(n_1-1, n_2), & n_1 = 1, 2, \dots, N_1, \\ \frac{\binom{n_1+n_2}{n_2}}{c(n_1, n_2-1)} \frac{d(n_1, n_2)}{d(n_1, n_2)} p_N(n_1, n_2-1), & n_2 = 0, 1, \dots, N_2, \end{cases}$$

$$\begin{aligned} n_1 &= 0, 1, \dots, N_1, \\ n_2 &= 1, 2, \dots, N_2, \end{aligned} \quad [4.45]$$

and

$$p_N(0,0) = \left[1 + \sum_{n_1=1}^{N_1} \frac{n_1^{-1} g(j)}{\prod_{j=0}^{h(j+1)}} + \sum_{n_2=1}^{N_2} \frac{n_2^{-1} r(j)}{\prod_{j=0}^{s(j+1)}} + \sum_{n_1=0}^{N_1-1} \sum_{n_2=1}^{N_2} \left[\frac{\binom{n_1+n_2+1}{n_1}}{\prod_{j=0}^{n_2-1} \frac{u(n_1, j)}{v(n_1, j+1)}} \right] \right]^{-1} \quad [4.46]$$

where the parameters $a(n_1, n_2)$, $b(n_1, n_2)$, $c(n_1, n_2)$ and $d(n_1, n_2)$ represents 'up' and 'down' flow coefficients (see Figure 25) given by

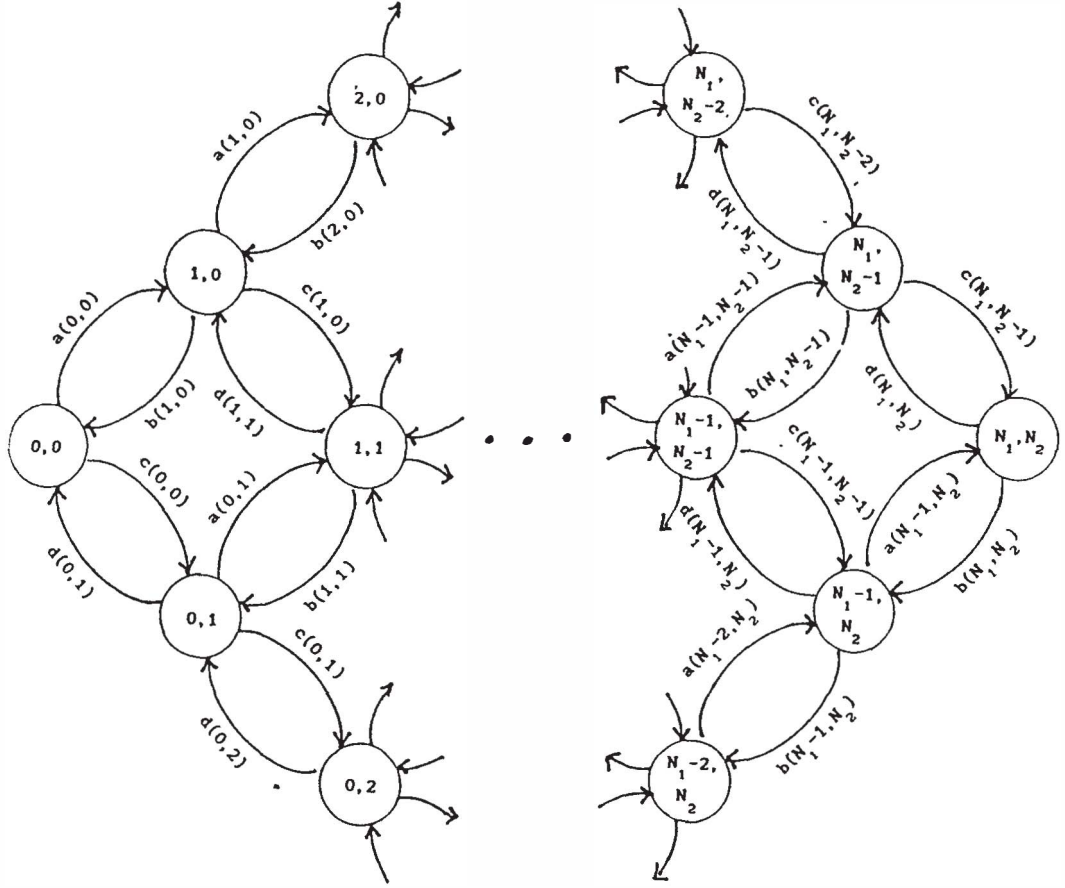


Figure 25 : Transition 'Up' and 'Down' Rate of a GE/GE/1/N with Two Classes of Users

$$a(n_1, n_2) = \begin{cases} \pi \lambda \tau_1 \tau_2, & n_1 = 0, n_2 = 0, \\ \pi e, & n_1 = 0, 1, \dots, N_1 - 2, \\ \frac{\pi e [f - (1 - \pi) e]}{\tau_1 \tau_2 (\mu - \lambda_2)}, & n_2 = 0, 1, \dots, N_2, \\ \frac{f \pi e (\mu - \lambda_1)}{\mu (f - \pi e)}, & n_1 = N_1 - 1, n_2 = 0, 1, \dots, N_2 - 1, \\ & n_1 = N_1 - 1, n_2 = N_2, \end{cases} \quad \left. \begin{matrix} n_1 + n_2 > 0, \\ n_1 = N_1 - 1, n_2 = N_2, \end{matrix} \right\}$$

[4.47]

$$b(n_1, n_2) = f \quad n_1 = 1, 2, \dots, N_1, \quad n_2 = 0, 1, \dots, N_2, \quad [4.48]$$

$$c(n_1, n_2) = \begin{cases} (1-\pi)\lambda\tau_1\tau_2, & n_1 = 0, n_2 = 0, \\ (1-\pi)e, & n_1 = 0, 1, \dots, N_1, \\ (1-\pi)e(f-\pi e) & n_2 = 0, 1, \dots, N_2-2, \end{cases} \left. \vphantom{\begin{matrix} (1-\pi)\lambda\tau_1\tau_2 \\ (1-\pi)e \\ (1-\pi)e(f-\pi e) \end{matrix}} \right\} n_1 + n_2 > 0, \\ \frac{\tau_1\tau_2(\mu-\lambda_1)}{f(1-\pi)e(\mu-\lambda_2)} & n_1 = 0, 1, \dots, N_1-1, n_2 = N_2-1, \\ \frac{\mu[f-(1-\pi)e]}{\mu[f-(1-\pi)e]} & n_1 = N_1, n_2 = N_2-1, \end{cases} \quad [4.49]$$

$$d(n_1, n_2) = f \quad n_1 = 0, 1, \dots, N_1, \quad n_2 = 1, 2, \dots, N_2, \quad [4.50]$$

and $g(n_1)$, $h(n_1)$, $r(n_2)$, $s(n_2)$, $u(n_1, n_2)$ and $v(n_1, n_2)$ are given by

$$g(n_1) = \begin{cases} \pi\lambda\tau_1\tau_2, & n_1 = 0, \\ \pi e, & n_1 = 1, 2, \dots, N_1-2, \\ \frac{\pi e[f-(1-\pi)e]}{\tau_1\tau_2(\mu-\lambda_2)}, & n_1 = N_1-1, \end{cases} \quad [4.51]$$

$$h(n_1) = s(n_2) = f, \quad n_1 = 1, 2, \dots, N_1, \quad n_2 = 1, 2, \dots, N_2, \quad [4.52]$$

$$r(n_2) = \begin{cases} (1-\pi)\lambda\tau_1\tau_2, & n_2 = 0, \\ (1-\pi)e, & n_2 = 1, 2, \dots, N_2-2, \\ (1-\pi)e(f-\pi e) & n_2 = N_2-1, \\ \frac{\tau_1\tau_2(\mu-\lambda_1)}{\tau_1\tau_2(\mu-\lambda_1)}, & \end{cases} \quad [4.53]$$

$$u(n_1, n_2) = \begin{cases} \left[\frac{\pi e}{f} \right]^n \left[\frac{f - (1-\pi)e}{\tau_1 \tau_2 (\mu - \lambda_2)} \right]^{h_1(n_1)} \frac{\lambda_1 \tau_1 \tau_2 (1-\pi)e}{f}, & n_1 = 0, \dots, N_1 - 1, \\ & n_2 = 0, \\ (1-\pi)e, & n_1 = 0, \dots, N_1 - 1, \\ & n_2 = 1, 2, \dots, N_2 - 2, \\ (1-\pi)e \left[\frac{f - \pi e}{\tau_1 \tau_2 (\mu - \lambda_1)} \right]^{h_2(n_1, n_2)} \left[\frac{f(\mu - \lambda_2)}{\mu[f - (1-\pi)e]} \right]^{h_2(n_1, n_2)}, & n_1 = 0, \dots, N_1 - 1, \\ & n_2 = N_2 - 1, \end{cases} \quad [4.54]$$

$$v(n_1, n_2) = f \quad \begin{matrix} n_1 = 0, \dots, N_1 - 1, \\ n_2 = 1, 2, \dots, N_2, \end{matrix} \quad [4.55]$$

$$h_1(n_1) = \begin{cases} 1, & \text{if } (n_1 = N_1 - 1 \text{ \& } n_2 = 0), \\ 0, & \text{otherwise.} \end{cases}$$

$$h_2(n_1, n_2) = \begin{cases} 1, & \text{if } (n_1 < N_1 - 1 \text{ \& } n_2 = N_2 - 1) \text{ or } (n_1 = N_1 - 1 \text{ \& } n_2 = N_2 - 1) \\ 0, & \text{otherwise.} \end{cases}$$

$$e = \mu \tau_2 (1 - \tau_1) + \lambda \tau_1, \quad f = \mu \tau_2 + \lambda \tau_1 (1 - \tau_2),$$

$$\lambda = \lambda_1 + \lambda_2, \quad \pi = \lambda_1 / \lambda, \quad \tau_1 = \frac{2}{(C_a^2 + 1)}, \quad \tau_2 = \frac{2}{(C_s^2 + 1)},$$

$$C_a^2 = \frac{\lambda_1 + \lambda_2}{\frac{\lambda_1}{(C_{\lambda_1}^2 + 1)} + \frac{\lambda_2}{(C_{\lambda_2}^2 + 1)}} - 1, \quad C_s^2 = C_{\mu}^2,$$

$$\langle N \rangle = \frac{\rho}{2} \left[1 + \frac{C_a^2 + \rho C_s^2}{1 - \rho} \right].$$

Note that Figure 25 is in a two dimensional form. So for three or more classes of users, the flow 'up' and 'down' rates will be in a three or more dimensional form. The mean throughput, mean queue length (any first order statistics) for the network model described above can easily be computed from the joint probabilities.

Extension to Load Dependent Queueing Systems

Consider a cyclic queueing model $GE(n_1+n_2)/GE(n_1+n_2)/1/N$ with two classes of users (i.e $K = 2$) as depicted in Figure 26. Centres 1 and 2 accept users at the rates $\lambda_1(N_1-n_1)$ and $\lambda_2(N_2-n_2)$ respectively which depends on the number of users n_1 and n_2 at centre 0. From [4.45] we can write $p_N(n_1, n_2)$ and $p_N(0, n_2)$ as

$$p_N(n_1, n_2) = \binom{n_1+n_2}{n_1} \prod_{j=0}^{n_1-1} \frac{a(j, n_2)}{b(j+1, n_2)} p_N(0, n_2), \quad [4.56]$$

and

$$p_N(0, n_2) = \prod_{j=0}^{n_2-1} \frac{c(0, j)}{d(0, j+1)} p_N(0, 0), \quad [4.57]$$

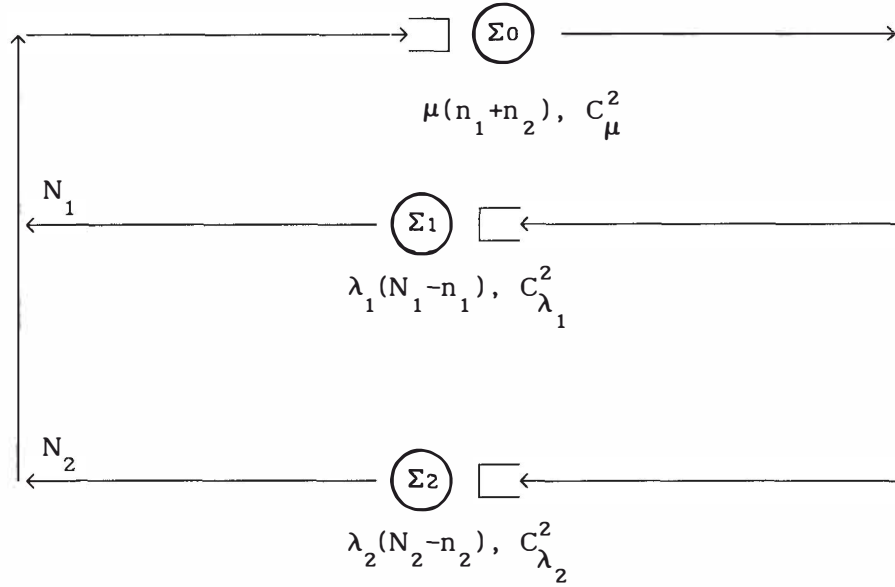


Figure 26 : Queueing Model for Two Classes of Users and Load Dependent Centres

By using [4.56] and [4.57] we can rewrite [4.45] and [4.46] as

$$p_N(n_1, n_2) = \binom{n_1+n_2}{n_1} \left[\prod_{j=0}^{n_1-1} \frac{a(j, n_2)}{b(j+1, n_2)} \right] \left[\prod_{j=0}^{n_2-1} \frac{c(0, j)}{d(0, j+1)} \right] p_N(0, 0), \quad [4.58]$$

and

$$p_N(0, 0) = \left[1 + \sum_{\substack{n_1=0 \\ (n_1+n_2>0)}}^{N_1} \sum_{n_2=0}^{N_2} \binom{n_1+n_2}{n_1} \left[\prod_{j=0}^{n_1-1} \frac{a(j, n_2)}{b(j+1, n_2)} \right] \left[\prod_{j=0}^{n_2-1} \frac{c(0, j)}{d(0, j+1)} \right] \right]^{-1}, \quad [4.59]$$

where $\prod_{j=0}^{-1} = 1$ and the parameters $a(n_1, n_2)$, $b(n_1, n_2)$, $c(n_1, n_2)$ and $d(n_1, n_2)$ are as given by [4.47]–[4.50] with λ , λ_1 and λ_2 given by $\lambda(N-n_1-n_2)$, $\lambda_1(N_1-n_1)$ and $\lambda_2(N_2-n_2)$ respectively.

Numerical Examples

In this section, numerical examples are presented to assess the credibility of the maximum entropy analytical approach and also to illustrate how critically the distributional form of interarrival and service times with known first two moments affects system performance in terms of primary measures such as throughput and mean queue length. The results are compared with those obtained by simulation. The simulation was based on the GE distribution representing general distributions with known first two moments. The simulation results quoted are produced using SIMSCRIPT II.5 (Russell, 1989). All measurement of time are in second. No comparison was made with results reported by Kouvatsos (1986b) since the number of user per class in the general queueing network studied by Kouvatsos (1986b) is not fixed. As mentioned earlier, the number of user per class in this study is fixed and no class switching are allowed.

The results for the network depicted in Figure 22 with $K = 2$ are presented in Figures 27–30 and for $K = 3$ are presented in Figures 31–34 respectively. The results for the network depicted in Figure 35 are shown in Figures 36–37. The fourth example is depicted in Figure 38 and the results for this example are presented in Figures 39–40. In the third and fourth examples, subnetworks 1 and 2 are reduced to single composite servers with load dependent service rates $\lambda_1(N_1 - n_1)$ and $\lambda_2(N_2 - n_2)$ respectively by using the generalised Norton's flow equivalent model reduction (Ku-Mahamud and Othman, 1990; 1993) or as presented in Chapter III. SIM represents the simulation done on the network in Figure 22 for the first two examples and the networks in

Figures 35 and 38 for the last two examples, while ME represents the results obtained using the proposed maximum entropy analytical approach.

The comparative study focuses on the throughput and mean queue length of centre 0 for the networks shown in Figures 22, 35 and 38. In the first example where $K = 2$, $\lambda_1 = 0.1$ (taken to be fixed for validation purposes), various values of λ_2 , C_μ^2 takes the value 1.0, $C_{\lambda_1}^2 = 0.5$, $C_{\lambda_2}^2 = 1.0$, $N_1 = 5$ and $N_2 = 10$. It is also noted that the GE type solution for the overall throughput as a function of the total offered load provide lower bounds (pessimistic) when $C_\mu^2 = 5.0$ and upper bounds (optimistic) when $C_\mu^2 = 0.5$ (see Figure 29). As expected, the overall mean queue length as a function of the total offered load

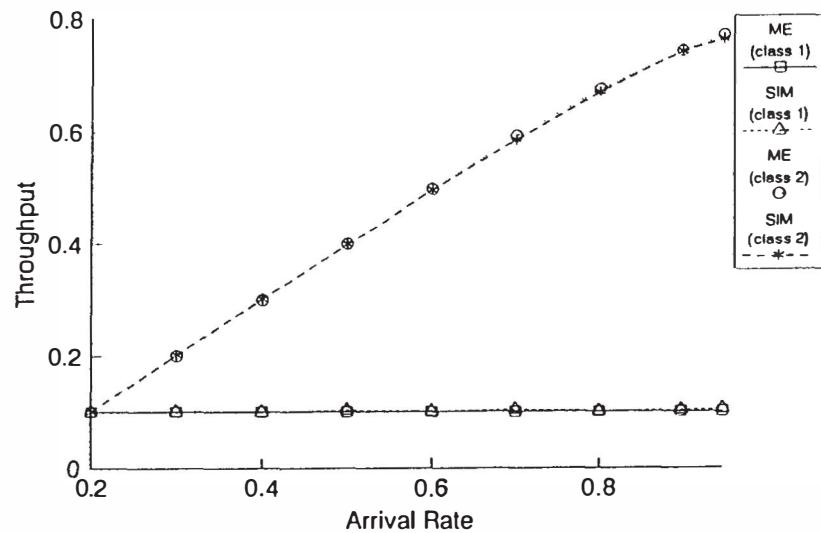


Figure 27 : Throughput of Class 1 and 2 Users versus Total Offered Load ($N_1 = 5$, $N_2 = 10$)

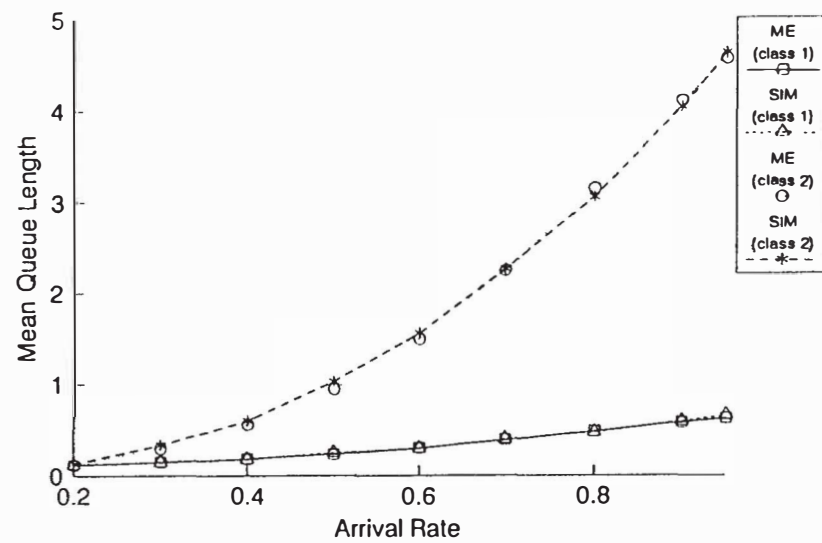


Figure 28 : Mean Queue Length of Class 1 and 2 Users versus Total Offered Load ($N_1 = 5$, $N_2 = 10$)

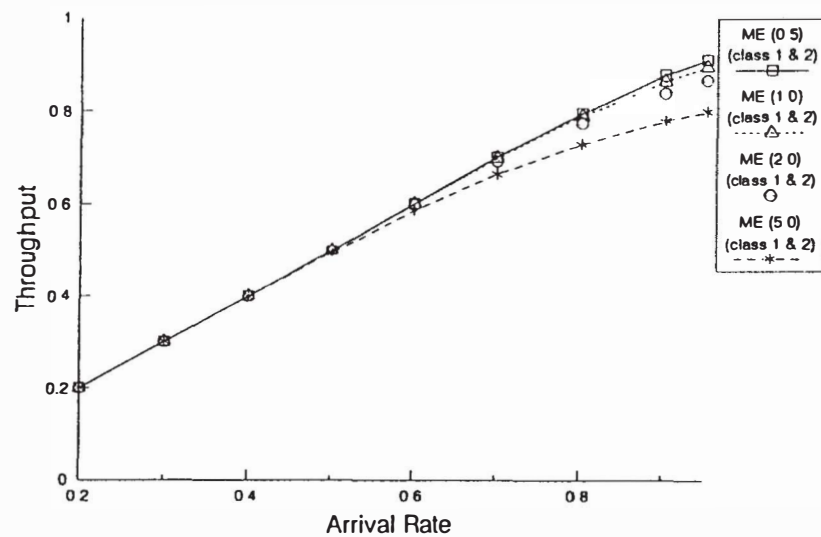


Figure 29 : Overall Throughput versus Total Offered Load ($N_1 = 5$, $N_2 = 10$)

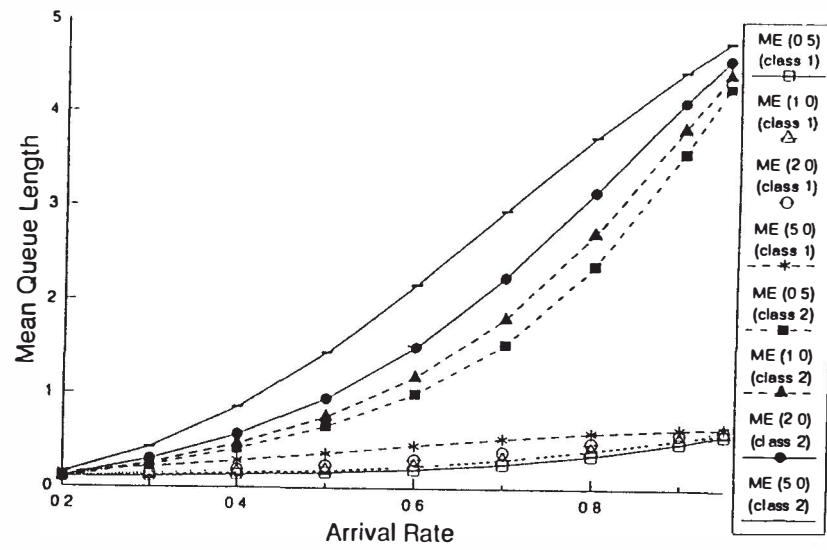


Figure 30 : Overall Mean Queue Length versus Total Offered Load
($N_1 = 5$, $N_2 = 10$)

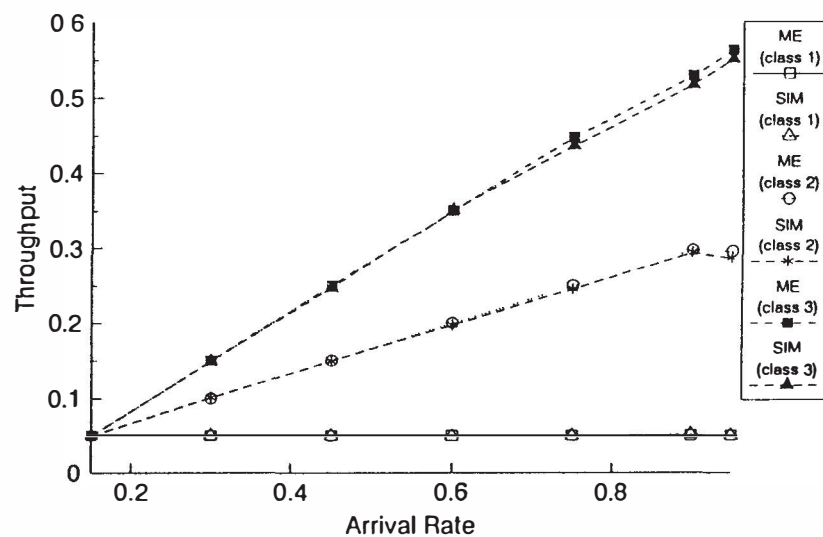


Figure 31 : Throughput of Class 1, 2 and 3 Users versus
Total Offered Load ($N_1 = N_2 = N_3 = 5$)

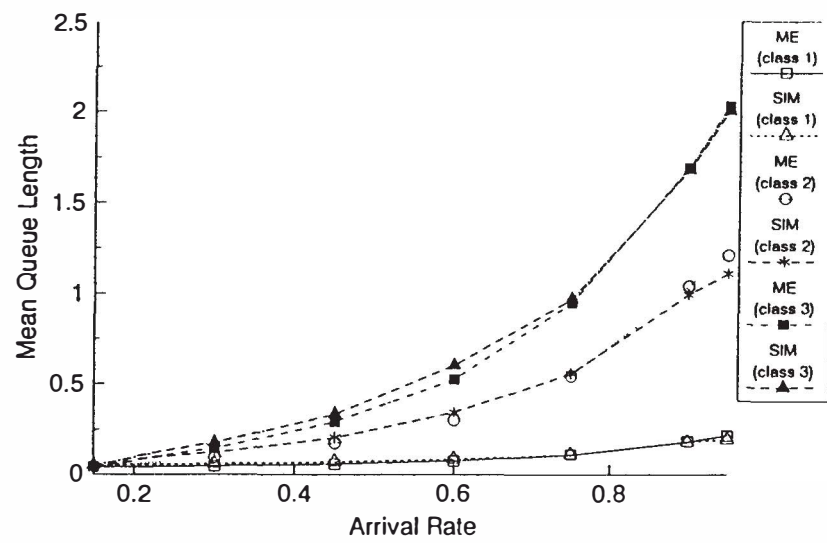


Figure 32 : Mean Queue Length of Class 1, 2 and 3 Users versus Total Offered Load ($N_1 = N_2 = N_3 = 5$)

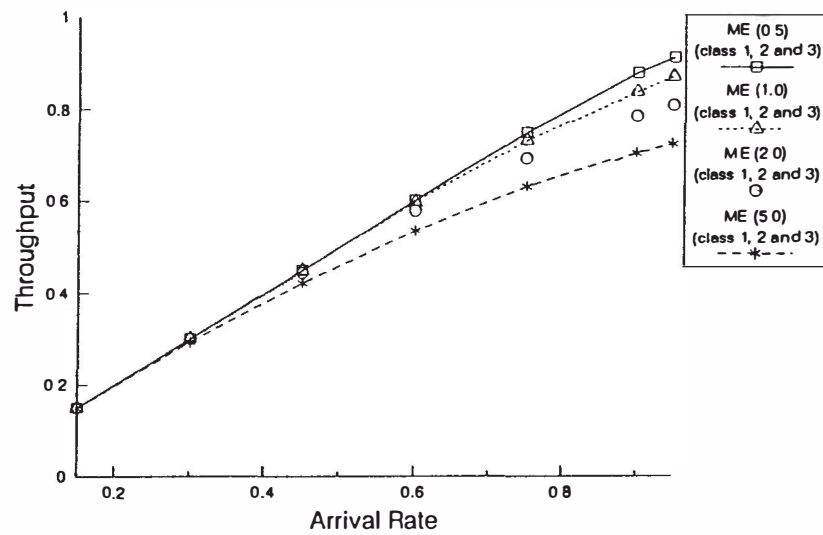


Figure 33 : Overall Throughput versus Total Offered Load ($N_1 = N_2 = N_3 = 5$)

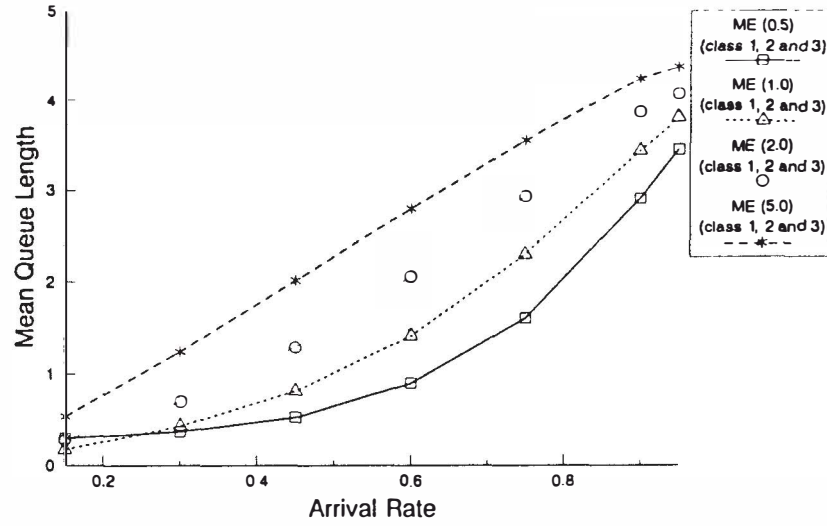


Figure 34 : Overall Mean Queue Length versus Total Offered Load
 $(N_1 = N_2 = N_3 = 5)$

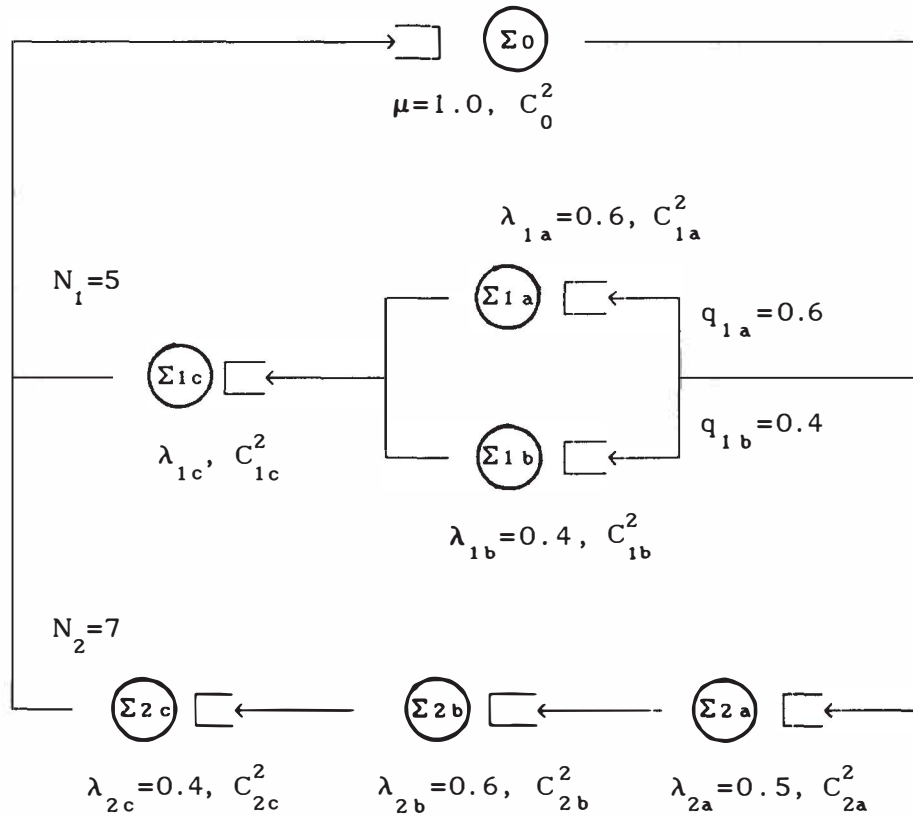


Figure 35 : Virtual Circuit Link Model for Two Classes of Users
 $(C_0^2 = C_{1a}^2 = C_{1b}^2 = C_{1c}^2 = C_{2a}^2 = C_{2b}^2 = C_{2c}^2 = 0.5)$

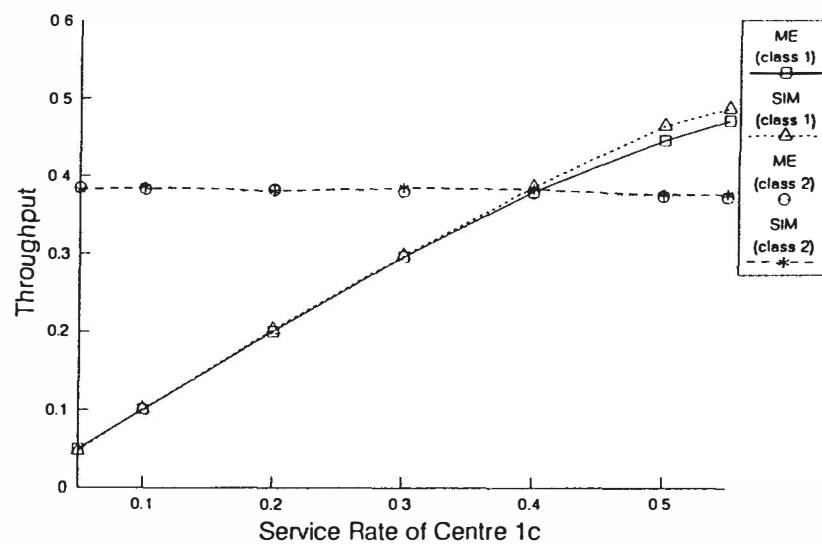


Figure 36 : Throughput of Class 1 and 2 Users versus Total Offered Load

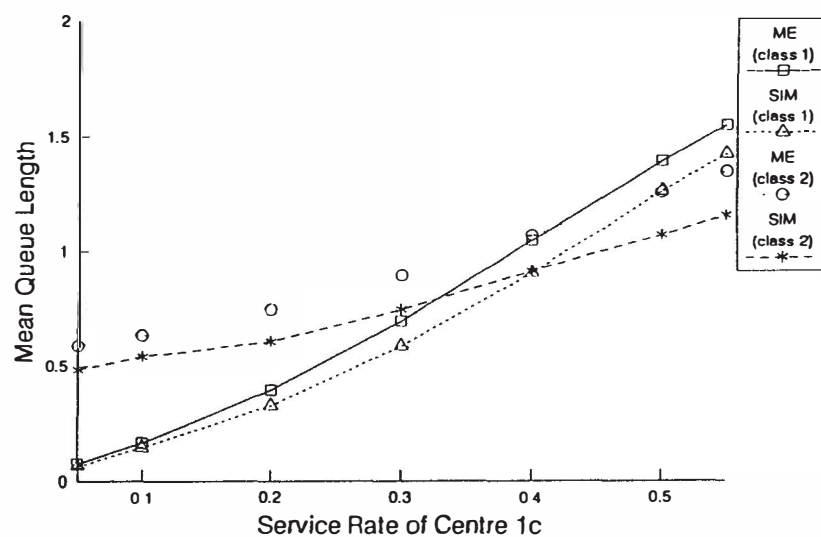


Figure 37 : Mean Queue Length of Class 1 and 2 Users versus Total Offered Load

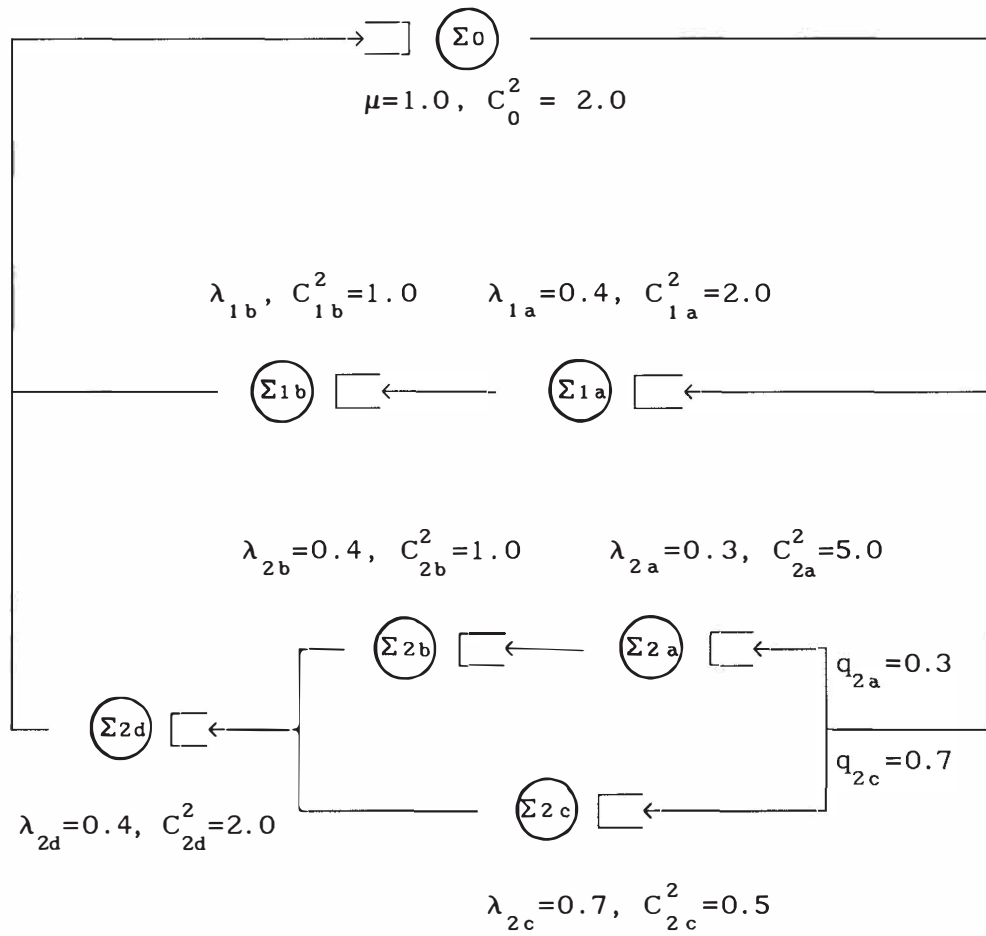


Figure 38 : A Queueing Model for Virtual Circuit Link with Two Classes of Users

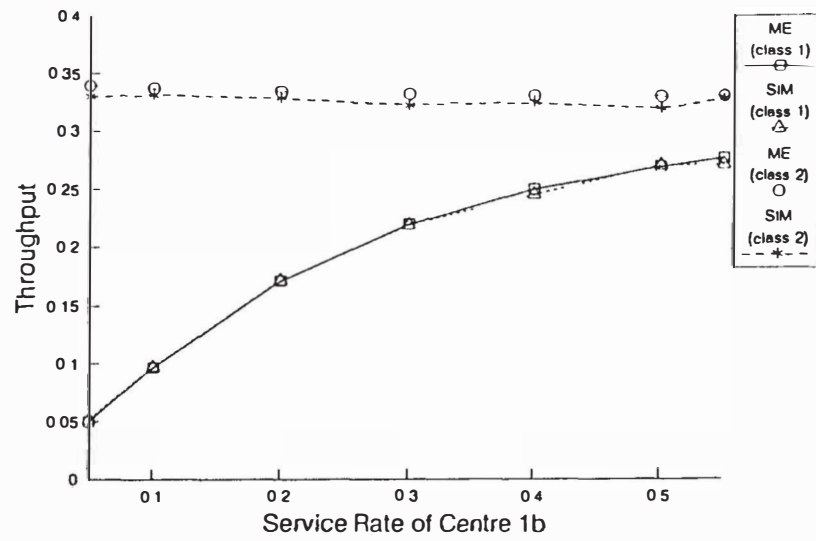


Figure 39 : Throughput of Class 1 and 2 Users versus Total Offered Load

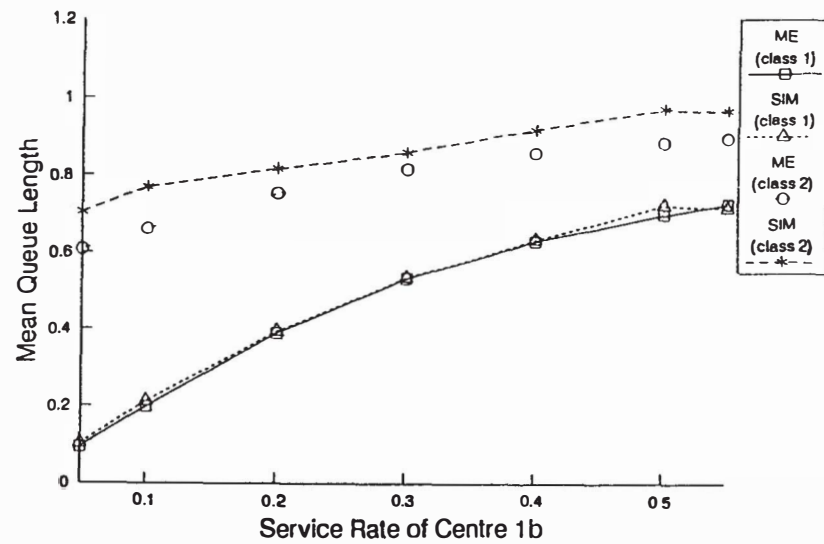


Figure 40 : Mean Queue Length of Class 1 and 2 Users versus Total Offered Load

provide lower bounds (pessimistic) when $C_{\mu}^2 = 0.5$ and upper bounds (optimistic) when $C_{\mu}^2 = 5.0$ (see Figure 30). The values for other parameters remain unchanged in this bounds tests. For the case where $K = 3$, $\lambda_1 = 0.05$, various values of λ_2 and λ_3 , $N_1 = N_2 = N_3 = 5$, and the squared coefficient of variation for all the centres take the same value (i.e. $C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = C_{\lambda_3}^2 = 0.5$). Again it is noted that the GE type solution for the overall throughput as a function of the total offered load provide lower bounds (pessimistic) when $C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = C_{\lambda_3}^2 = 5.0$ and upper bounds (optimistic) when $C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = C_{\lambda_3}^2 = 0.5$ (see Figure 33). As expected, the overall mean queue length as a function of the total offered load provide lower bounds (pessimistic) when the squared coefficient of the centres are 0.5 and upper bounds (optimistic) when the squared coefficient of variation of the centres are 5.0 (see Figure 34). Again, the values for other parameters remain unchanged in this bounds tests. In the third example (see Figures 36 and 37) the throughput and mean queue length are plotted against various values of λ_{1c} while in the fourth example (see Figures 39 and 40), the throughput are plotted against various values of λ_{1b} . μ is taken to be 1 in all the examples. It can be generally observed that the ME results are comparable to those obtained by SIM under all traffic condition. Note that in the case of exponential distributed network (i.e. $C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = C_{\lambda_3}^2 = 1$), the results obtained from the maximum entropy approach is the same as reported in classical queueing theory.

Summary

The principle of maximum entropy has been used to derive a new product form solution for the approximate analysis of a mixed class (K closed) queueing network at equilibrium with infinite capacities, single server and FCFS service discipline. Numerical evaluation results demonstrated the credibility of the maximum entropy approach, and also provide useful insights into how critically the system performance is affected by the distributional form of the service times. Moreover favourable comparisons in terms of throughput and mean queue length are made with simulation.

CHAPTER V

MULTISERVER GENERAL QUEUEING NETWORKS WITH MULTICLASS USERS

Introduction

In this chapter, a new approximate analytic solution for the joint queue length distribution of a multiserver general closed queueing network with multiclass users is proposed. The multiserver general closed queueing network with multiclass users can be used to model several practical situations such as the machine repairman problem (Allen, 1978) and multiprogramming computer systems under heavy traffic conditions where it is assumed that a constant number of users (i.e. tasks) circulating in the systems (Takacs, 1962; Kleinrock, 1975b). The multiserver general closed queueing network with multiclass users can also be used to model the decentralised optimal flow control of computer communication networks with multiple transmission links and multiple controllers (Hsiao and Lazar, 1991).

This chapter is organised as follows. A detailed description of the multiserver general closed queueing network with multiclass users is presented followed by the derivation of the analytic maximum entropy formulation of the multiserver general closed queueing network with multiclass users. This derivation is based on the known heuristic arguments and earlier results regarding the study of the stable single server general closed queueing network with multiclass users. The one-step maximum entropy recursions solution is then presented followed by the solution to load dependent servers. Numerical examples

will be provided to test the credibility of the new solution followed by a summary of this chapter.

Description of the Network

The general queueing network we now consider is as depicted in Figure 41. This network consists of $K+1$ infinite queueing centres and K classes of users. Centre 0 consists of c ($c \geq 2$) parallel homogeneous servers under FCFS service discipline and is modelled as a general server with fixed service rate μ , constant coefficient of variation C_μ . Users belonging to either class queue up in a common queue to wait for service. After receiving service at centre 0, users of class i , $i = 1, 2, \dots, K$, proceed to centre i for their next service. Centre i accept users of class i at a rate λ_i and constant coefficient of variation of C_{λ_i} on a FCFS basis. The maximum number of class i users is N_i and the maximum number of users in the queueing network model of Figure 41 is $N (= \sum_{i=1}^K N_i)$.

Maximum Entropy for Multiserver General Queueing System with Multiclass Users

In this section, the maximum entropy analytical approach used in solving a restricted (Kouvatsos and Almond, 1988b) multiclass (i.e. K closed) and multiserver general queueing network model as depicted in Figure 45 is presented. The lower queues (i.e. centres 1, 2, ..., K) are single servers with mean service rate λ_i and coefficient of variation C_{λ_i} . These servers generate a restricted arrival process to centre 0 which is a c homogeneous multiserver queue with mean service

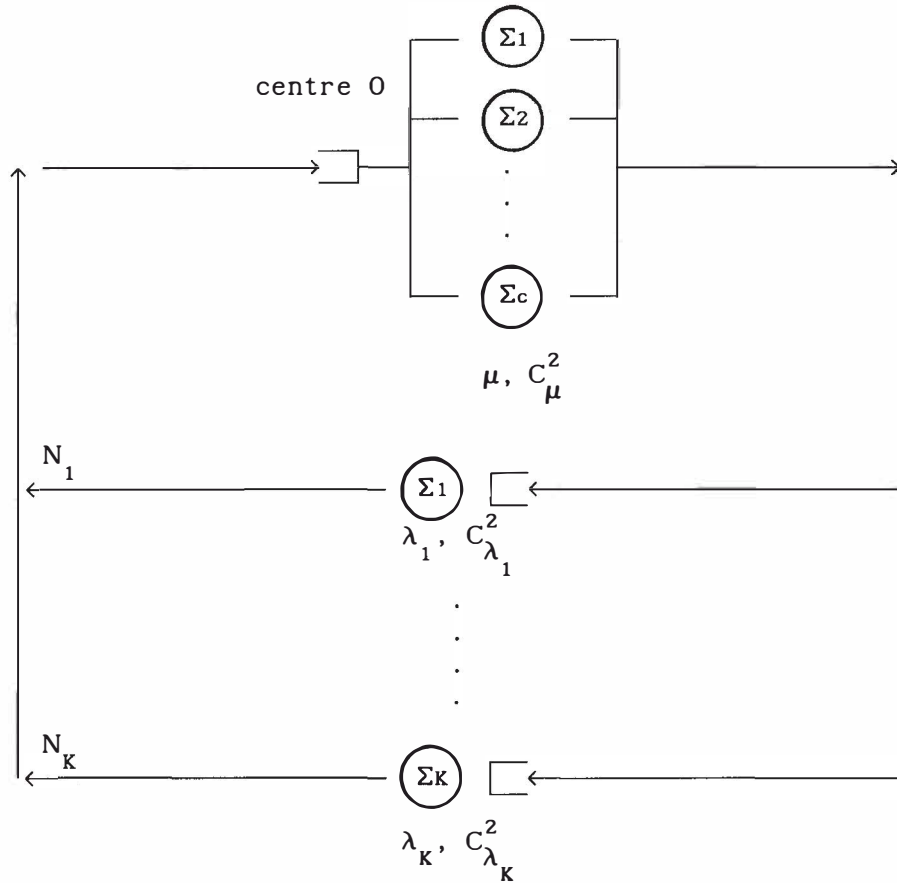


Figure 41 : Queueing Model for Multiclass Users with Single Server and Multiserver Queueing Centres

rate μ and coefficient of variation C_{μ}^2 . Note that since the routing depends on the class of users, the general closed network model is an extended Baskett *et al.* (1975) network and can be analysed by using the maximum entropy approximation approach. Let $\underline{n} = (n_1, n_2, \dots, n_k)$ be the state of centre 0 where $n_i, n_i = 0, 1, \dots, N_i$, and N_i is the total number of users of class i . Furthermore let $N = \sum_{i=1}^k N_i$ be the total number of users in the queueing system. Centre 0 can be seen as a restricted multiserver general closed queueing system with multiclass users denoted by $G/G/c/N$. In order to reduce the number of

known information needed in the analysis, the following assumptions are made.

- i. Restricted multiserver queueing systems behave like restricted single server queueing system with parameters λ , C_λ , $\hat{\mu}$, C_μ if the number of users in the system is greater than the number of servers (i.e. $n \geq c$, $\hat{\mu} = c\mu$, $\lambda = \sum_{i=1}^K \lambda_i$, C_λ is the coefficient of variation of the arrivals from centres 1, 2, ..., K).
- ii. c servers queueing systems have identical transition weights up to the $c-2$ states with $c-1$ servers queueing systems (i.e. for $c \geq 2$).

Note that assumption (i) is equal to a busy $G/G/1/N$ queue with service time $1/c\mu$. In fact this assumption had been used by many analysts simplifying their stochastic behaviour of single class users and multiserver models (Kelton and Law, 1985). Assumption (ii) is heuristically true for any stable multiserver queueing system under heavy traffic and exact for the case of limiting (i.e. $N \rightarrow \infty$) $M/M/c/N$, $M/G/c/N$ (Allen, 1978; Kleinrock, 1975a), $GE/M/c/N$ (Takacs, 1962), and $GE/GE/c/N$ (Kouvatsos, 1986b).

Based on the above assumptions, we can easily formulate the maximum entropy solution for the restricted general multiclass and c multiserver queueing system (i.e. $G/G/c/N$) provided that we have complete information about the analysis of $G/G/c-1/N$ and the corresponding $G/G/1/N$ queueing system. Let us now consider the $G/G/c/N$ queue and assumed that the parameters λ_i , μ , $C_{\lambda_i}^2$ and C_μ^2 , form a basic set of a *priori* knowledge. Suppose all that is known about the

steady state joint probability $p_N(\underline{n})$, $\underline{n} = \{n_1, n_2, \dots, n_K\}$, at centre 0 are the following mean value constraints:

i. the normalisation condition,

$$\sum_{n_1=0}^{N_1} \dots \sum_{n_K=0}^{N_K} p_N(\underline{n}) = 1. \quad [5.1]$$

ii. the probabilities $P_N(\hat{n} \geq i) = U(i)$, $0 < U(i) < 1$, $i = 1, 2, \dots, c$,

$$\sum_{n_1=0}^{N_1} \dots \sum_{n_K=0}^{N_K} h_i(\underline{n}) p_N(\underline{n}) = U(i), \quad [5.2]$$

$$h_i(\underline{n}) = \begin{cases} 0, & \text{if } n_1 + n_2 + \dots + n_K < i, \\ 1, & \text{otherwise,} \end{cases}$$

where $\hat{n}$ is a random variable describing the queue length.

iii. the mean queue length per class $\langle n_i \rangle_{N_i}$, $\rho_i \leq \langle n_i \rangle_{N_i} < N_i$, where

$$\rho_i = \lambda_i / \mu \text{ and } \langle N_c \rangle_N = \sum_{i=1}^K \langle n_i \rangle_{N_i},$$

$$\sum_{n_1=0}^{N_1} \dots \sum_{n_K=0}^{N_K} n_i p_N(\underline{n}) = \langle n_i \rangle_{N_i}, \quad i = 1, 2, \dots, K, \quad [5.3]$$

iv. the flow balance,

$$\begin{aligned}
 & \left(\sum_{i=1}^K \lambda_i \right) \sum_{n_1=0}^{N_1-1} \sum_{n_2=0}^{\infty} \dots \sum_{n_K=0}^{\infty} p_N(\underline{n}) + \left(\sum_{i=2}^K \lambda_i \right) \sum_{n_2=0}^{\infty} \dots \sum_{n_K=0}^{\infty} p_N(N_1, n_2, \dots, n_K) \\
 & = \mu \left[\sum_{n_1=0}^{N_1} \sum_{n_2=0}^{\infty} \dots \sum_{n_K=0}^{\infty} b(\underline{n}) p_N(\underline{n}) \right], \quad [5.4]
 \end{aligned}$$

for $n_1 = N_1$, $n_i < N_i$ ($i = 2, 3, \dots, K$),

.

.

.

.

.

$$\left(\sum_{i=1}^K \lambda_i \right) \sum_{n_1=0}^{N_1-1} \dots \sum_{n_K=0}^{N_K-1} p_N(\underline{n}) + \lambda_1 \sum_{n_1=0}^{N_1-1} p_N(n_1, N_2, \dots, N_K) + \dots +$$

$$\lambda_K \sum_{n_K=0}^{N_K} p_N(N_1, \dots, N_{K-1}, n_K) = \mu \left[\sum_{n_1=0}^{N_1} \dots \sum_{n_K=0}^{N_K} b(\underline{n}) p_N(\underline{n}) \right], \quad [5.5]$$

for $n_i = N_i$ ($i = 1, 2, \dots, K$),

$$b(\underline{n}) = \begin{cases} n_1 + n_2 & n_1 + n_2 < c, \\ c & \text{otherwise.} \end{cases}$$

Note that there will be a total of $\sum_{k=0}^{K-1} 2^k$ flow balance constraints.

This corresponds to a total of $\sum_{k=0}^{K-1} 2^k$ ways the state of centre 0 {i.e.

$\underline{n}$ contains at least one $n_i = N_i$. The above statistics may be known numerically via system measurements during finite observations periods

or can be determined via known analytic formulae (if available) which are usually expressed in terms of moments of the interarrival and service time distributions.

The formulation of the maximum entropy for the restricted multiclass G/G/c/N queue will begin from the multiclass G/G/1 queue as follows. For simplicity, let us consider the case of $K = 2$ classes of users but the results could easily be extended to $K > 2$ classes of users as will be shown later. From Chapter IV, the maximum entropy solution of a G/G/1 queue with two classes of users subject to constraints [5.1]–[5.3] is given by

$$p(\underline{n}) = p(0) \frac{(n_1 + n_2)!}{n_1! n_2!} g_1^{h_1(\underline{n})} x_1^{n_1(1)} x_2^{n_2(1)} \quad [5.6]$$

$$h_1(\underline{n}) = \begin{cases} 1, & \text{if } n_1 + n_2 \geq 1, \\ 0, & \text{otherwise,} \end{cases}$$

where $p(0)$ is obtained from normalisation and

$$g_1 = \frac{[1 - x(1)] \rho(1)}{[1 - \rho(1)] x(1)}, \quad [5.7]$$

$$x(1) = \frac{\langle N_1 \rangle - \rho(1)}{\langle N_1 \rangle} = x_1^{(1)} + x_2^{(1)}, \quad [5.8]$$

$$x_i^{(1)} = \frac{\langle n_i(1) \rangle (1 - x(1)) x(1)}{\rho(1)}, \quad i = 1, 2, \quad [5.9]$$

$$\rho(1) = \rho_1^{(1)} + \rho_2^{(1)}, \quad \rho_i^{(1)} = \lambda_i / \mu,$$

and $\langle N_1 \rangle$ is a known mean number of users in a general single server queue. Let us now consider the G/G/2 queue with one step-transition weight as depicted in Figure 42. The corresponding entropy solution can be written as

$$p(\underline{n}) = p(\underline{0}) \begin{pmatrix} n_1 + n_2 \\ n_1 \end{pmatrix} g_1^{h_1(\underline{n})} g_2^{h_2(\underline{n})} x_1^{n_1(2)} x_2^{n_2(2)} \quad [5.10]$$

subject to constraints:

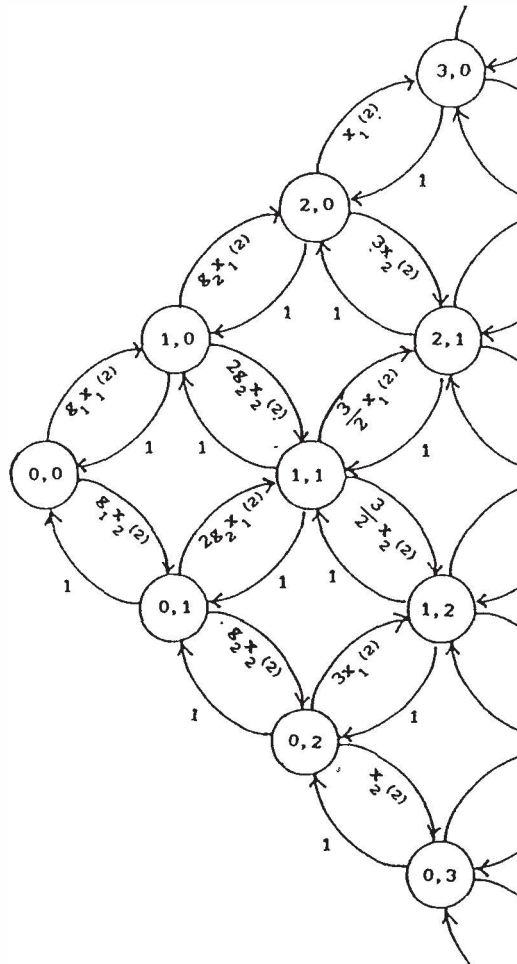


Figure 42 : Transition Weight of a G/G/2 Queue with Two Classes of Users

i. normalisation

$$\sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} p(\underline{n}) = 1, \quad [5.11]$$

ii. the probabilities $P_N(\hat{n} \geq i) = U(i)$, $0 < U(i) < 1$, $i = 1, 2$,

$$\sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} h_i(\underline{n}) p(\underline{n}) = U(i), \quad [5.12]$$

iii. the mean queue length per class $\langle n_i(2) \rangle$, $\rho_i(2) \leq \langle n_i(2) \rangle < N_i$,

$$\text{where } \rho_i(2) = \frac{\lambda_i}{2\mu}, \quad \langle N_2 \rangle = \langle n_1(2) \rangle + \langle n_2(2) \rangle,$$

$$\sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} n_i p(\underline{n}) = \langle n_i(2) \rangle, \quad i = 1, 2. \quad [5.13]$$

iv. the flow balance

$$(\lambda_1 + \lambda_2) \sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} p(\underline{n}) = \mu \left[\sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} b(\underline{n}) p(\underline{n}) \right] \quad [5.14]$$

$$x(2) = \frac{\langle N_2 \rangle - \rho(2)}{\langle N_2 \rangle}, \quad \text{where } \langle N_2 \rangle \text{ due to assumption (i) is given by } \langle N_2 \rangle$$

$= \langle \hat{N}_1 \rangle$ is the mean queue length formula for a single server queue with adjusted parameter (i.e $\rho(1) \leftarrow \rho(2)$) is known for any available exact or approximate explicit formula. It can be verified that

$$p(\underline{0}) = \frac{x(2) - (\lambda_1 + \lambda_2)/\mu + [1 - x(2)] \langle N_2 \rangle}{x(2)}, \quad [5.15]$$

$$g_1 = \frac{(\lambda_1 + \lambda_2)/\mu [2 - x(2)] - 2 [1 - x(2)] \langle N_2 \rangle}{x^2(2) p(0)}, \quad [5.16]$$

$$g_2 = \frac{[1 - x(2)]^2 [\langle N_2 \rangle - p(0) g_1 x(2)]}{x^2(2)[2 - x(2)] p(0) g_1}, \quad [5.17]$$

By knowing the solution of the G/G/2 queue, we can solve the G/G/3 queue where g_1 for both queues has a common form due to assumption (ii). The one-step transition weight of the G/G/3 queue can be depicted in Figure 43 and the corresponding entropy solution is given by

$$p(\underline{n}) = p(0) \binom{n_1 + n_2}{n_1} g_1^{h_1(\underline{n})} g_2^{h_2(\underline{n})} g_3^{h_3(\underline{n})} x_1^{n_1(2)} x_2^{n_2(2)}, \quad [5.18]$$

subject to constraints:

i. normalisation

$$\sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} p(\underline{n}) = 1, \quad [5.19]$$

ii. the probabilities $P_N(\hat{n} \geq i) = U(i)$, $0 < U(i) < 1$, $i = 1, 2, 3$,

$$\sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} h_i(\underline{n}) p(\underline{n}) = U(i), \quad [5.20]$$

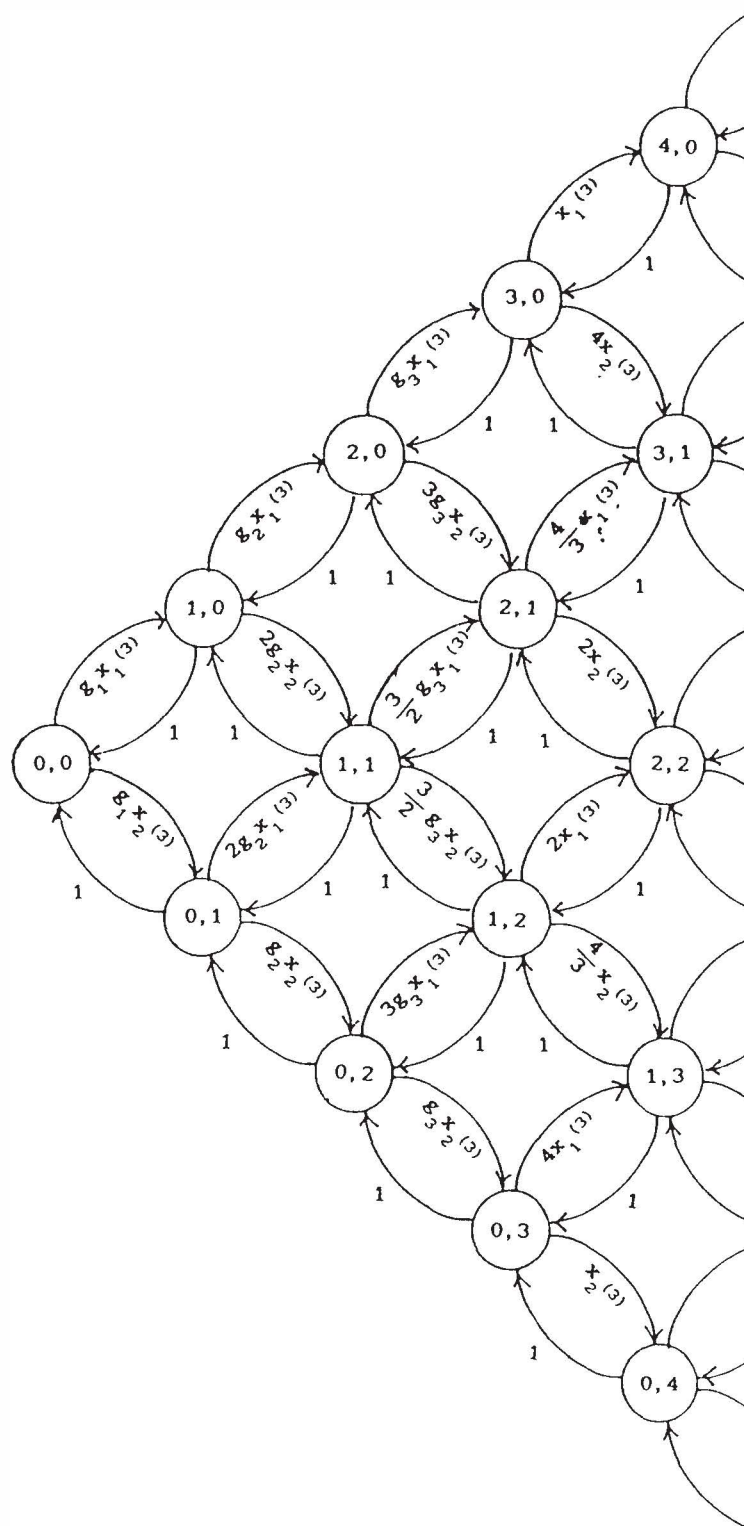


Figure 43 : Transition Weight of a G/G/3 Queue with Two Classes of Users

- iii. the mean queue length per class $\langle n_i(3) \rangle$, $\rho_i(3) \leq \langle n_i(3) \rangle < N_i$,
 where $\rho_i(3) = \frac{\lambda_i}{3\mu}$, $\langle N3 \rangle = \langle n_1(3) \rangle + \langle n_2(3) \rangle$,

$$\sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} n_i p(\underline{n}) = \langle n_i(3) \rangle, \quad i = 1, 2. \quad [5.21]$$

- iv. the flow balance

$$(\lambda_1 + \lambda_2) \sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} p(\underline{n}) = \mu \left[\sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} b(\underline{n}) p(\underline{n}) \right], \quad [5.22]$$

where $\langle N3 \rangle$ is known by similar assumption as in G/G/2 queue. Since g_1 is assumed to be known from G/G/2 queue and $x(3)$ is obtained due to assumption (i), then $p(\underline{0})$, g_2 and g_3 can be estimated by satisfying the above constraints and are given by

$$p(\underline{0}) = \frac{2x(3) - (\lambda_1 + \lambda_2)/\mu + [1 - x(3)] \langle N3 \rangle}{x(3) [2 + g_1 x(3)]}, \quad [5.23]$$

$$g_2 = \frac{(\lambda_1 + \lambda_2)/\mu [3 - 2x(3)] - 3[1 - x(3)] \langle N3 \rangle - p(\underline{0}) g_1 x^2(3)}{2g_1 x^3(3) p(\underline{0})}, \quad [5.24]$$

$$g_3 = \frac{[1 - x(3)]^2 \left[\langle N3 \rangle - [1 + 2g_2 x(3)] p(\underline{0}) g_1 x(3) \right]}{x^3(3) [3 - 2x(3)] p(\underline{0}) g_1 g_2}, \quad [5.25]$$

In general, the above assumption can be extended to G/G/c queue by assuming the transition weight g_i ($i = 1, 2, \dots, c-2$) are known from

$G/G/c-1$ queue and $x(c)$ is obtained due to assumption (i). This can be done as follows. The one-step transition weight diagram for $G/G/c$ queue can be drawn as shown in Figure 44 and the corresponding entropy solution is given by

$$p(\underline{n}) = p(\underline{0}) \begin{pmatrix} n_1 + n_2 \\ n_1 \end{pmatrix} \left[\prod_{i=1}^c \frac{h_i(\underline{n})}{g_i} \right] x_1^{n_1(c)} x_2^{n_2(c)}, \quad [5.26]$$

subject to constraints:

i. normalisation

$$\sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} p(\underline{n}) = 1, \quad [5.27]$$

ii. the probabilities $P_N(\hat{n} \geq i) = U(i)$, $0 < U(i) < 1$, $i = 1, 2, \dots, c$,

$$\sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} h_i(\underline{n}) p(\underline{n}) = U(i), \quad [5.28]$$

iii. the mean queue length per class $\langle n_{i(c)} \rangle$, $\rho_{i(c)} \leq \langle n_{i(c)} \rangle < N_i$,
where $\rho_{i(c)} = \frac{\lambda_i}{c\mu}$, $\langle N_c \rangle = \langle n_{1(c)} \rangle + \langle n_{2(c)} \rangle$,

$$\sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} n_i p(\underline{n}) = \langle n_{i(c)} \rangle, \quad i = 1, 2. \quad [5.29]$$

iv. the flow balance

$$(\lambda_1 + \lambda_2) \sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} p(\underline{n}) = \mu \left[\sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} b(\underline{n}) p(\underline{n}) \right] \quad [5.30]$$

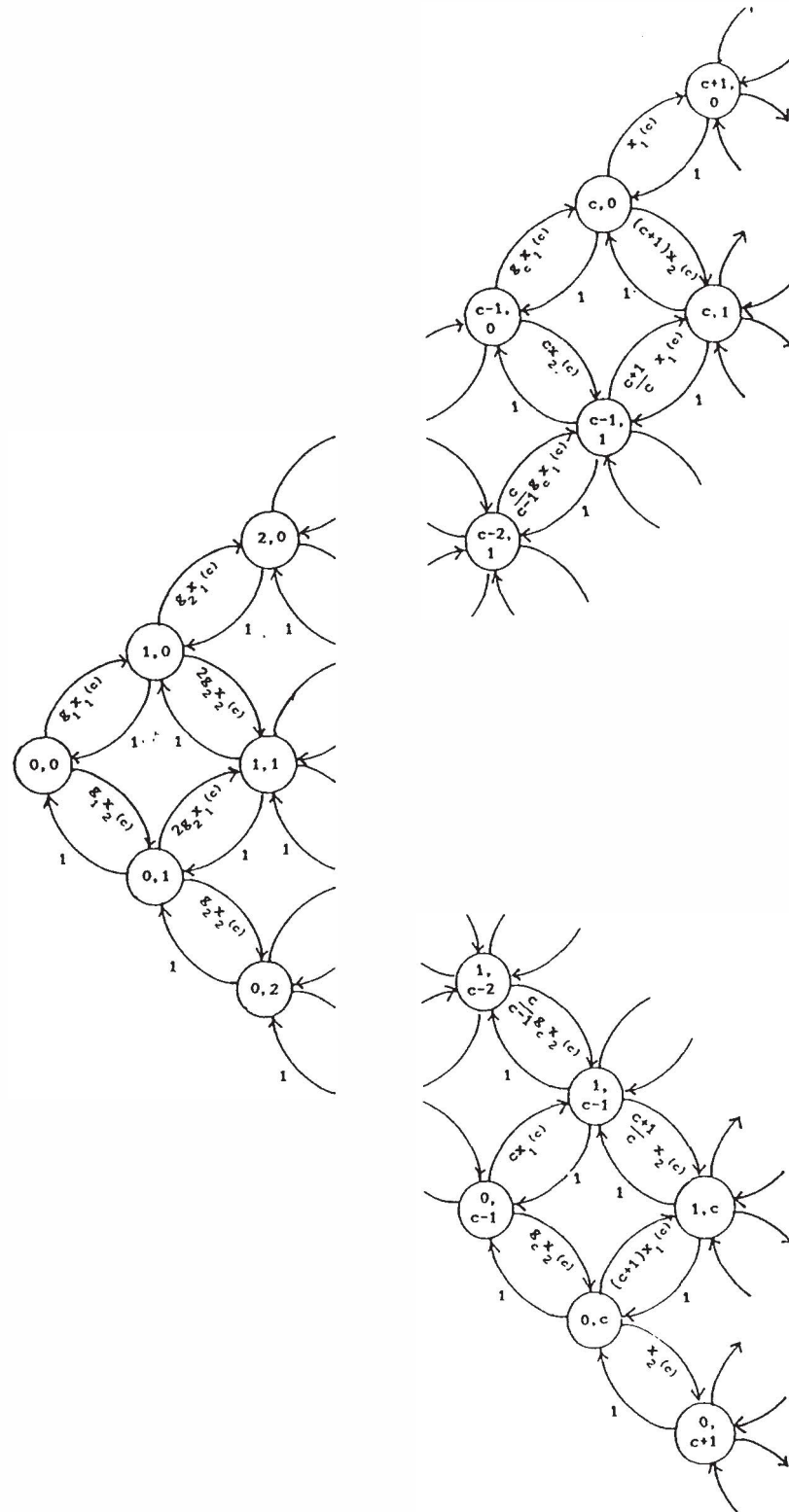


Figure 44 : Transition Weight of a G/G/c Queue with Two Classes of Users

where similarly $\langle N_c \rangle$ is known. It can be verified that

$$p(\underline{0}) = \frac{(c-1)x(c) - (\lambda_1 + \lambda_2)/\mu + [1 - x(c)] \langle N_c \rangle}{x(c) \left[(c-1) + (c-2) \sum_{j=1}^{c-2} [x^j(c) \prod_{i=1}^j g_i] \right]}, \quad [5.31]$$

$$g_{c-1} = \frac{\frac{(\lambda_1 + \lambda_2)}{\mu} [c - (c-1)x(c)] - c[1 - x(c)] \langle N_c \rangle - p(\underline{0}) \sum_{j=0}^{c-2} [j x^{j+1}(c) \prod_{i=1}^j g_i]}{p(\underline{0}) x^c(c) (c-1) \prod_{i=1}^{c-2} g_i}, \quad [5.32]$$

$$g_c = \frac{[1 - x(c)]^2 \left[\langle N_c \rangle - p(\underline{0}) \sum_{j=1}^{c-1} [j x^j(c) \prod_{i=1}^j g_i] \right]}{x^c(c) [c - (c-1)x(c)] p(\underline{0}) \prod_{i=1}^{c-1} g_i}, \quad [5.33]$$

where $\prod_{i=1}^0 g_i = 1$. From the above queue length distribution of the G/G/c queue, we can formulate the entropy solution for the restricted G/G/c/N queue by assuming that all g_i ($i = 1, 2, \dots, c$) and $x(c)$ are independent of N provided that $c \ll N$. By analogue, the resultant transition weight for this queue is as depicted in Figure 45 and the corresponding entropy solution is given by

$$p_N(\underline{n}) = p_N(\underline{0}) \binom{n_1 + n_2}{n_1} \left[\prod_{i=1}^c g_i^{h_i(\underline{n})} \right] \left[\prod_{i=1}^2 x_i^{n_i}(c) \right] \left[\prod_{i=1}^3 y_i^{f_i(\underline{n})}(c) \right] \quad [5.34]$$

$$f_1(\underline{n}) = \begin{cases} 1 & \text{if } n_1 = N_1 \text{ \& } n_2 \leq N_2, \\ 0 & \text{otherwise,} \end{cases}$$

$$f_2(\underline{n}) = \begin{cases} 1 & n_1 \leq N_1 \text{ \& } n_2 = N_2, \\ 0 & \text{otherwise,} \end{cases}$$

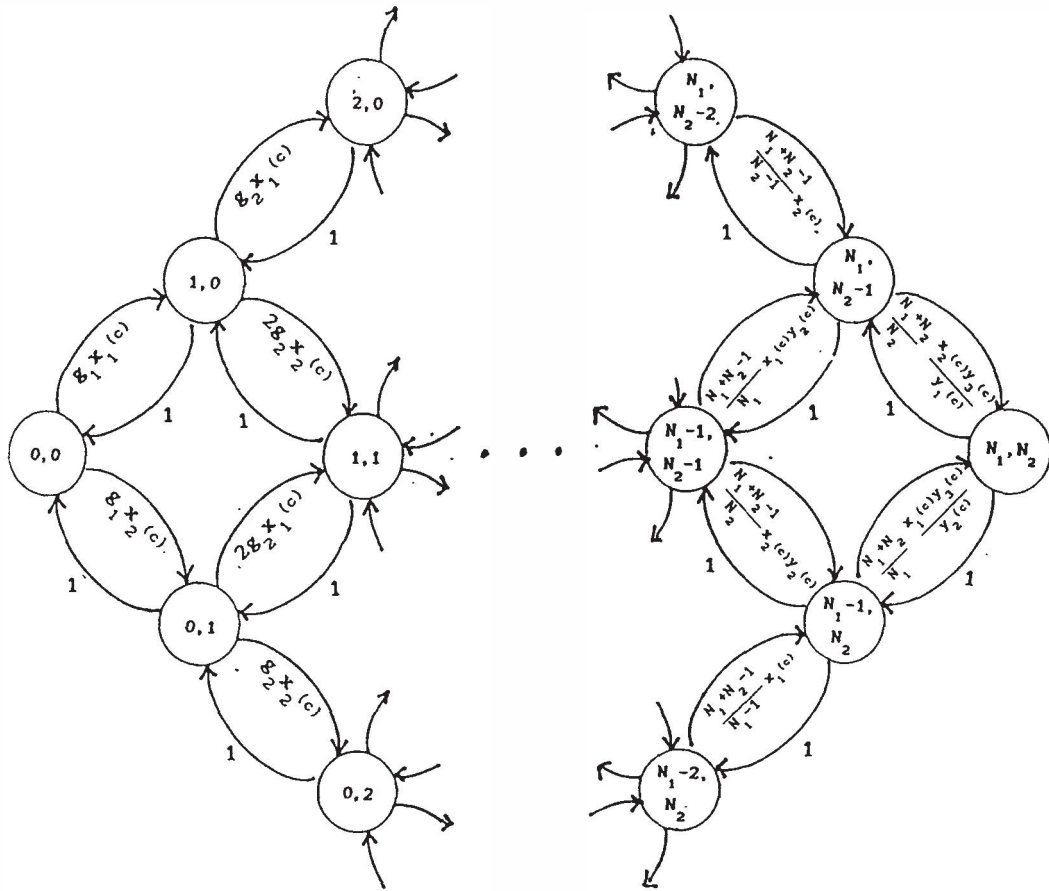


Figure 45 : Transition Weight of a G/G/c/N Queue with Two Classes of Users

$$f_3(\underline{n}) = \begin{cases} 1 & n_1 = N_1 \text{ \& \; } n_2 = N_2, \\ 0 & \text{otherwise.} \end{cases}$$

where $p_N(0)$ is given by

$$p_N(0) = \frac{(c-1)x(c) - (\lambda_1 + \lambda_2)/\mu + [1 - x(c)] \langle Nc \rangle_N}{x(c) \left[(c-1) + (c-2) \sum_{j=1}^{c-2} [x^j(c) \prod_{i=1}^j g_i] \right]}, \quad [5.35]$$

and

$$x(c) = \frac{\langle Nc \rangle - \rho(c)}{\langle Nc \rangle} = x_1(c) + x_2(c), \quad [5.36]$$

$$x_i(c) = \frac{\langle n_1(c) \rangle (1 - x(c)) x(c)}{\rho(c)}, \quad i = 1, 2, \quad [5.37]$$

$$g_{c-1} = \frac{\frac{(\lambda_1 + \lambda_2)}{\mu} [c - (c-1)x(c)] - c[1 - x(c)] \langle Nc \rangle_N - p_N(0) \sum_{j=0}^{c-2} [j x^{j+1}(c) \prod_{i=1}^j g_i]}{p_N(0) x^c(c) (c-1) \prod_{i=1}^{c-2} g_i}, \quad [5.38]$$

$$g_c = \frac{[1 - x(c)]^2 \left[\langle Nc \rangle_N - p_N(0) \sum_{j=1}^{c-1} [j x^j(c) \prod_{i=1}^j g_i] \right]}{x^c(c) [c - (c-1)x(c)] p_N(0) \prod_{i=1}^{c-1} g_i}, \quad [5.39]$$

where $\prod_{i=1}^0 g_i = 1$. The parameters $y_1(c)$, $y_2(c)$ and $y_3(c)$ can be obtained by satisfying the flow balance conditions [5.4]–[5.5] and are given by

$$y_1(c) = \frac{(1 - \rho(c))(1 - x_2(c))}{(1 - x(c))(1 - \rho_2(c))}, \quad [5.40]$$

$$y_2(c) = \frac{(1 - \rho(c))(1 - x_1(c))}{(1 - x(c))(1 - \rho_1(c))}, \quad [5.41]$$

$$y_3(c) = \frac{1 - \rho(c)}{1 - x(c)}. \quad [5.42]$$

The above solution can easily be extended to the $G/G/c/N$ ($c \geq 2$ parallel

servers, $K > 2$ classes) queue and the joint queue length distribution is given by

$$p_{N-}(n) = \frac{1}{Z(p)} \frac{\left[\sum_{i=1}^K n_i \right]!}{\left[\prod_{i=1}^K n_i! \right]} \left[\prod_{i=1}^K \frac{c_i^{h_i(n)}}{g_i} \right] \left[\prod_{i=1}^K x_i^{n_i} \right] y_1^{f(1)(c)} \dots y_t^{f(t)(c)},$$

$$n_i = 0, 1, \dots, N_i$$
[5.43]

$$f(j) = \begin{cases} 1, & \text{if } j \text{ corresponds to one of the ways that centre } 0 \\ & \text{contains at least one } n_i = N_i, i = 1, 2, \dots, K, \\ 0, & \text{otherwise,} \end{cases}$$

$$j = 1, 2, \dots, t = \sum_{k=0}^{K-1} 2^k,$$

subject to constraints [5.1]–[5.5]. $p_{N-}(0)$ can be obtained from normalisation condition and is given by

$$p_{N-}(0) = \left[\sum_{n_1=0}^{N_1} \dots \sum_{n_K=0}^{N_K} \frac{\left[\sum_{i=1}^K n_i \right]!}{\left[\prod_{i=1}^K n_i! \right]} \left[\prod_{i=1}^K \frac{c_i^{h_i(n)}}{g_i} \right] \left[\prod_{i=1}^K x_i^{n_i} \right] y_1^{f(1)(c)} \dots y_t^{f(t)(c)} \right]^{-1}$$
[5.44]

and

$$x_i(c) = \frac{\langle n_i \rangle (1 - x(c)) x(c)}{\rho(c)}, \quad i = 1, 2, \dots, K,$$
[5.45]

$$x(c) = \frac{\sum_{i=1}^K \langle n_i \rangle - \rho(c)}{\sum_{i=1}^K \langle n_i \rangle}, \quad x_1(c) = \pi x(c), \quad \pi = \frac{\lambda_1}{\lambda_1 + \dots + \lambda_K}, \quad [5.46]$$

$$y_j(c) = \frac{(1 - \rho(c)) \left[1 - \sum_{\substack{k=1 \\ k \neq \text{finite capacity node}}}^K x_k(c) \right]}{(1 - x(c)) \left[1 - \sum_{\substack{k=1 \\ k \neq \text{finite capacity node}}}^K \rho_k(c) \right]}, \quad [5.47]$$

One-Step Maximum Entropy Recursions

From the joint maximum entropy solution [5.43], the following one-step maximum entropy recursions are established.

$$p_N(1,0,\dots,0) = p_N(0) g_1 x_1(c), \quad [5.48]$$

$$\vdots$$

$$p_N(0,0,\dots,1) = p_N(0) g_1 x_K(c), \quad [5.49]$$

$$\vdots$$

$$p_N(n_1,0,\dots,0) = x_1(c) g_{n_1}^{h(n)} y_1^{f(1)}(c) p_N(n_1-1,0,\dots,0), \quad n_1 = 2, 3, \dots, N_K, \quad [5.50]$$

$$\begin{aligned}
 & \vdots \\
 & \vdots \\
 & \vdots \\
 & p_N(0, \dots, 0, n_K) = x_K^{(c)} g_{n_K}^{h(\underline{n})} y_t^{f(t)} p_N(0, \dots, 0, n_K - 1), \quad n_K = 2, 3, \dots, N_K,
 \end{aligned} \tag{5.51}$$

$$h(\underline{n}) = \begin{cases} 1 & \text{if } (n_1 + n_2 \leq c), \\ 0 & \text{otherwise,} \end{cases}$$

$$p_N(\underline{n}) = \begin{cases} \frac{(n_1 + \dots + n_K)}{n_1} x_1^{(c)} g_{n_1 + \dots + n_K}^{h(\underline{n})} Z(y) p(n_1 - 1, n_2, \dots, n_K), \\ \vdots \\ \frac{(n_1 + \dots + n_K)}{n_K} x_K^{(c)} g_{n_1 + \dots + n_K}^{h(\underline{n})} Z(y) p(n_1, \dots, n_{K-1}, n_K - 1) \end{cases} \tag{5.52}$$

where $Z(y)$ represents the various combination of the Langrangian coefficients, y , that will occur when $n_i = N_i$, $i = 1, 2, \dots, K$. Recursions [5.48]–[5.52] enable analyst to study more general operational or stochastic K resource queues with finite buffer, N , for which operational Birth–Death transitions (Buzen and Denning, 1980) or classical queueing theory (Kleinrock, 1975a; Allen, 1978) no longer effectively apply. Due to the complexity in writing down the one-step maximum entropy recursion solutions when using either the bulk or the limiting interpretation of the GE distribution (Kouvatsos, 1988a), let us now for simplicity, consider for the case $K = 2$. [5.48]–[5.52] can now be written in the form given by

$$p_N(n_1, n_2) = \begin{cases} \frac{(n_1+n_2)}{n_1} x_1^{(c)} g_{n_1+n_2}^{h(n)} y_1^{f_1(n)} \left[\frac{y_3^{(c)}}{y_2^{(c)}} \right]^{f_3(n)} p_N(n_1-1, n_2), \\ \frac{(n_1+n_2)}{n_2} x_2^{(c)} g_{n_1+n_2}^{h(n)} y_2^{f_2(n)} \left[\frac{y_3^{(c)}}{y_1^{(c)}} \right]^{f_3(n)} p_N(n_1, n_2-1), \end{cases}$$

$$\begin{aligned} n_1 &= 0, 1, \dots, N_1, \\ n_2 &= 0, 1, \dots, N_2. \end{aligned} \quad [5.53]$$

It can be shown that by using either the bulk or the limiting interpretation of the GE distribution, [5.53] can be written in the form given by

$$p_N(n_1, n_2) = \begin{cases} \frac{(n_1+n_2)}{n_1} \frac{a(n_1-1, n_2)}{b(n_1, n_2)} p_N(n_1-1, n_2), & n_1 = 1, 2, \dots, N_1, \\ \frac{(n_1+n_2)}{n_2} \frac{c(n_1, n_2-1)}{d(n_1, n_2)} p_N(n_1, n_2-1), & n_2 = 0, 1, \dots, N_2, \end{cases}$$

$$\begin{aligned} n_1 &= 0, 1, \dots, N_1, \\ n_2 &= 1, 2, \dots, N_2, \end{aligned} \quad [5.54]$$

and

$$p_N(0,0) = \left[1 + \sum_{n_1=1}^{N_1-1} \left[\prod_{j=0}^{n_1-1} \frac{g(j)}{h(j+1)} \right] + \sum_{n_2=1}^{N_2-1} \left[\prod_{j=0}^{n_2-1} \frac{r(j)}{s(j+1)} \right] + \right.$$

$$\left. \sum_{n_1=0}^{N_1-1} \sum_{n_2=1}^{N_2} \left[\left(\frac{n_1+n_2+1}{n_1} \right) \prod_{j=0}^{n_2-1} \frac{u(n_1, j)}{v(n_1, j+1)} \right] \right]^{-1}, \quad [5.55]$$

where the parameters $a(n_1, n_2)$, $b(n_1, n_2)$, $c(n_1, n_2)$ and $d(n_1, n_2)$ represents 'up' and 'down' flow coefficients given by (see Figure 46)

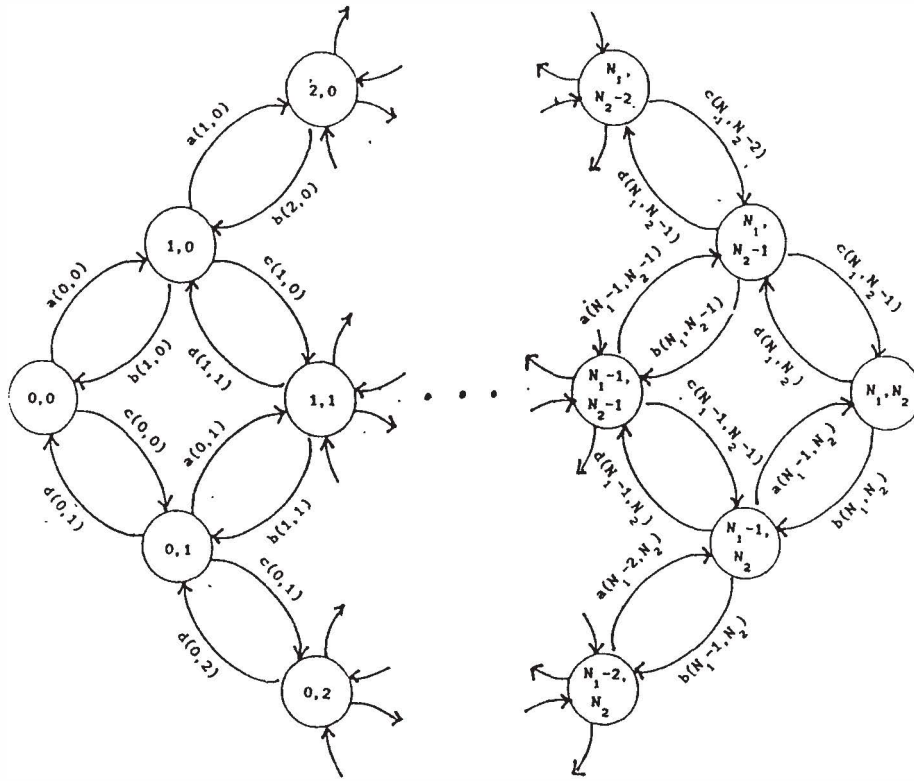


Figure 46 : Transition 'Up' and 'Down' Rate Solution of a GE/GE/c/N Queue with Two Classes of Users

$$a(n_1, n_2) = \begin{cases} \pi \lambda \tau_1, & n_1 + n_2 = 0, \\ \pi t, & n_1 + n_2 = 1, 2, \dots, c-2, \\ \pi \tau_2 e, & n_1 + n_2 = c-1, \\ \pi q, & n_1 = 0, 1, \dots, N_1-2, \\ & n_2 = 0, 1, \dots, N_2, \end{cases} \left. \begin{array}{l} \\ \\ \\ \end{array} \right\} n_1 + n_2 \geq c, \\ \begin{cases} \frac{\pi q [f - (1-\pi)q]}{\tau_1 \tau_2 (c\mu - \lambda_2)}, & n_1 = N_1-1, \\ & n_2 = 0, 1, \dots, N_2-1, \end{cases} \left. \begin{array}{l} \\ \end{array} \right\} n_1 + n_2 \geq c, \\ \begin{cases} \frac{f \pi q (c\mu - \lambda_1)}{c\mu (f - \pi q)}, & n_1 = N_1-1, \\ & n_2 = N_2, \end{cases} \left. \begin{array}{l} \\ \end{array} \right\} n_1 + n_2 \geq c,$$

[5.56]

$$b(n_1, n_2) = \begin{cases} (n_1 + n_2)\mu(\tau_1 + \tau_2 - \tau_1\tau_2) & n_1 + n_2 < c, \\ f & n_1 + n_2 \geq c, \end{cases} \quad [5.57]$$

$$c(n_1, n_2) = \begin{cases} (1-\pi)\lambda\tau_1, & n_1 + n_2 = 0, \\ (1-\pi)t, & n_1 = n_2 = 1, 2, \dots, c-2, \\ (1-\pi)\tau_2 e, & n_1 + n_2 = c-1, \\ (1-\pi)q, & \left. \begin{array}{l} n_1 = 0, 1, \dots, N_1, \\ n_2 = 0, 1, \dots, N_2-2, \end{array} \right\} n_1 + n_2 \geq c, \\ \frac{(1-\pi)q(f-\pi q)}{\tau_1\tau_2(c\mu-\lambda_1)}, & \left. \begin{array}{l} n_1 = 0, 1, \dots, N_1-1, \\ n_2 = N_2-1, \end{array} \right\} n_1 + n_2 \geq c, \\ \frac{f(1-\pi)q(c\mu-\lambda_2)}{c\mu[f-(1-\pi)q]}, & \left. \begin{array}{l} n_1 = N_1, \\ n_2 = N_2-1, \end{array} \right\} n_1 + n_2 \geq c, \end{cases} \quad [5.58]$$

$$d(n_1, n_2) = \begin{cases} (n_1 + n_2)\mu(\tau_1 + \tau_2 - \tau_1\tau_2) & n_1 + n_2 < c, \\ f & n_1 + n_2 \geq c, \end{cases} \quad [5.59]$$

and $g(n_1)$, $h(n_1)$, $r(n_2)$, $s(n_2)$, $u(n_1, n_2)$ and $v(n_1, n_2)$ are given by

$$g(n_1) = \begin{cases} \pi\lambda\tau_1, & n_1 = 0, \\ \pi t, & n_1 = 1, 2, \dots, c-2, \\ \pi\tau_2 e, & n_1 = c-1, \\ \pi q, & n_1 = c, c+1, \dots, N_1-2, \\ \frac{\pi q[f-(1-\pi)q]}{\tau_1\tau_2(c\mu-\lambda_2)}, & n_1 = N_1-1, \end{cases} \quad [5.60]$$

$$h(n_1) = s(n_2) = \begin{cases} (n_1 + n_2)\mu(\tau_1 + \tau_2 - \tau_1\tau_2) & n_1 + n_2 < c, \\ f & n_1 + n_2 \geq c, \end{cases} \quad [5.61]$$

$$r(n_2) = \begin{cases} (1-\pi)\lambda\tau_1, & n_2 = 0, \\ (1-\pi)t, & n_2 = 1, 2, \dots, c-2, \\ (1-\pi)\tau_2 e, & n_2 = c-1, \\ (1-\pi)q, & n_2 = c, c+1, \dots, N_2-2, \\ \frac{(1-\pi)q(f-\pi q)}{\tau_1\tau_2(c\mu-\lambda_1)}, & n_2 = N_2-1, \end{cases} \quad [5.62]$$

$$u(n_1, n_2) = \begin{cases} \lambda_2\tau_1[\pi t]^{h_1}[\pi\tau_2 e]^{h_2}[\pi q]^{h_3}\left[\frac{f-(1-\pi)q}{\tau_1\tau_2(c\mu-\lambda_2)}\right]^{h_4}, & n_1 = 0, 1, \dots, N_1-1, \\ & n_2 = 0, \\ h_5, & n_1 = 0, 1, \dots, N_1-1, \\ & n_2 = 1, 2, \dots, N_2-1, \end{cases} \quad [5.63]$$

$$v(n_1, n_2) = \begin{cases} \left[(n_1+n_2)\mu(\tau_1+\tau_2-\tau_1\tau_2)\right]^{h_6} f^{h_7}, & n_1 = 0, 1, \dots, N_1-1, \\ & n_2 = 1, \\ \left[(n_1+n_2)\mu(\tau_1+\tau_2-\tau_1\tau_2)\right]^{h_8} f^{h_9}, & n_1 = 0, 1, \dots, N_1-1, \\ & n_2 = 2, 3, \dots, N_2, \end{cases} \quad [5.64]$$

$$h_1 = \begin{cases} 0, & \text{if } c \leq 2, \\ n_1 + 1, & \text{if } (c \geq 3) \text{ \& } (c-n > 3), \\ c-2, & \text{if } (c \geq 3) \text{ \& } (c-n \leq 3), \end{cases}$$

$$h_2 = \begin{cases} 0, & \text{if } c-n > 2, \\ 1, & \text{if } c-n \leq 2, \end{cases}$$

$$h_3 = \begin{cases} 0, & \text{if } c-n > 1, \\ n-c+2, & \text{if } c-n \leq 1, \end{cases}$$

$$h_4 = \begin{cases} 0, & \text{if } n < N_1 - 1, \\ 1, & \text{if } n = N_1 - 1, \end{cases}$$

$$h_5 = \begin{cases} (1-\pi)t, & c-n_1-n_2 \geq 3, \\ (1-\pi)\tau_2 e, & c-n_1-n_2 = 2, \\ (1-\pi)q, & \left. \begin{array}{l} n_1 = 0, 1, \dots, N_1 - 1, \\ n_2 = 0, 1, \dots, N_2 - 2, \end{array} \right\} c-n_1-n_2 \leq 1, \\ \frac{(1-\pi)q(f-\pi q)}{\tau_1 \tau_2 (c\mu - \lambda_1)}, & \left. \begin{array}{l} n_1 = 0, 1, \dots, N_1 - 2, \\ n_2 = N_2 - 1, \end{array} \right\} c-n_1-n_2 \leq 1, \\ \frac{f(1-\pi)q(c\mu - \lambda_2)}{c\mu[f - (1-\pi)q]}, & \left. \begin{array}{l} n_1 = N_1 - 1, \\ n_2 = N_2 - 1, \end{array} \right\} c-n_1-n_2 \leq 1, \end{cases}$$

$$h_6 = \begin{cases} n_1 + n_2 + 1, & \text{if } n_1 + n_2 < c - 1, \\ c - 1, & \text{if } n_1 + n_2 \geq c - 1, \end{cases}$$

$$h_7 = \begin{cases} 0, & \text{if } n_1 + n_2 < c - 1, \\ n_1 + n_2 - c + 2, & \text{if } n_1 + n_2 \geq c - 1, \end{cases}$$

$$h_8 = \begin{cases} 1, & \text{if } n_1 + n_2 < c - 1, \\ 0, & \text{if } n_1 + n_2 \geq c - 1, \end{cases}$$

$$h_9 = \begin{cases} 1, & \text{if } n_1 + n_2 \geq c - 1, \\ 0, & \text{if } n_1 + n_2 < c - 1, \end{cases}$$

$$e = (c-1)\mu\tau_2(1 - \tau_1) + \lambda\tau_1, \quad q = c\mu\tau_2(1 - \tau_1) + \lambda\tau_1,$$

$$t = (n_1 + n_2)\mu\tau_2(1 - \tau_1) + \lambda\tau_1, \quad f = c\mu\tau_2 + \lambda\tau_1(1 - \tau_2),$$

$$\lambda = \lambda_1 + \lambda_2, \quad \pi = \lambda_1/\lambda, \quad \tau_1 = \frac{2}{(C_a^2 + 1)}, \quad \tau_2 = \frac{2}{(C_s^2 + 1)},$$

$$C_a^2 = \frac{\frac{\lambda_1 + \lambda_2}{\lambda_1} - 1}{\frac{1}{(C_{\lambda_1}^2 + 1)} + \frac{\lambda_2}{\lambda_1}} - 1, \text{ and } C_s^2 = C_{\mu}^2.$$

The mean throughput, mean queue length (i.e. any first order statistics) for the network model described above can easily be computed from the joint probabilities.

Extension to Load Dependent Queueing Systems

Consider a cyclic queueing model $GE(n_1+n_2)/GE(n_1+n_2)/c/N$ (i.e. $c \geq 2$, parallel servers) with two classes of users (i.e $K = 2$) as depicted in Figure 47. Centres 1 and 2 accept users at the rates $\lambda_1(N_1-n_1)$ and $\lambda_2(N_2-n_2)$ respectively which depends on the number of users n_1 and n_2 at centre 0 which accepts users at the rate $\mu(n_1+n_2)$. From [5.54] and [5.55], we can write $p_N(n_1, n_2)$ and $p_N(0, n_2)$ as

$$p_N(n_1, n_2) = \binom{n_1+n_2}{n_1} \prod_{j=0}^{n_1-1} \frac{a(j, n_2)}{b(j+1, n_2)} p_N(0, n_2), \quad [5.65]$$

and

$$p_N(0, n_2) = \prod_{j=0}^{n_2-1} \frac{c(0, j)}{b(0, j+1)} p_N(0, 0), \quad [5.66]$$

By using [5.65] and [5.66] we can rewrite [5.34] as

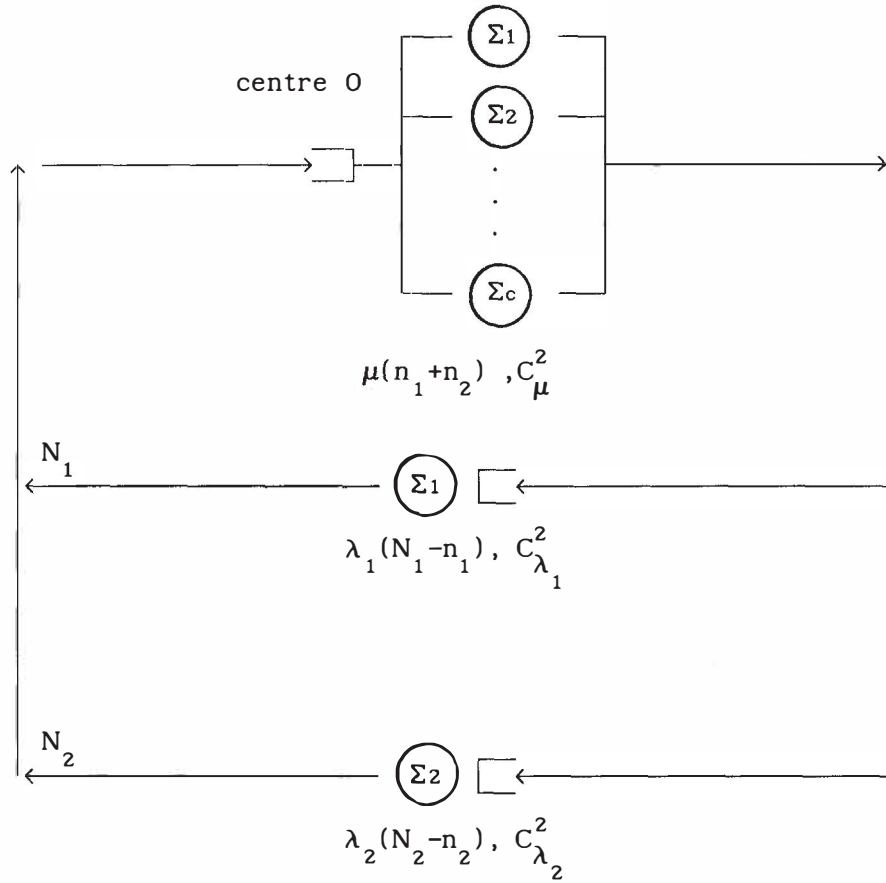


Figure 47 : Queueing Model for Two Classes of Users and Load Dependent Centres

$$p_N(n_1, n_2) = \binom{n_1+n_2}{n_1} \prod_{j=0}^{n_1-1} \frac{a(j, n_2)}{b(j+1, n_2)} \prod_{j=0}^{n_2-1} \frac{c(0, j)}{b(0, j+1)} p_N(0, 0), \quad [5.67]$$

and from normalisation condition, $p_N(0, 0)$ is given by

$$p_N(0, 0) = \left[1 + \sum_{\substack{n_1=0 \\ (n_1+n_2>0)}}^{N_1} \sum_{n_2=0}^{N_2} \binom{n_1+n_2}{n_1} \left[\prod_{j=0}^{n_1-1} \frac{a(j, n_2)}{b(j+1, n_2)} \right] \left[\prod_{j=0}^{n_2-1} \frac{c(0, j)}{b(0, j+1)} \right] \right]^{-1} \quad [5.68]$$

where the parameters $a(n_1, n_2)$, $b(n_1, n_2)$, $c(n_1, n_2)$ and $d(n_1, n_2)$ are as given by [5.56]–[5.59] with λ , λ_1 , λ_2 and μ given by $\lambda(N-n_1-n_2)$, $\lambda_1(N_1-n_1)$, $\lambda_2(N_2-n_2)$ and $\mu(n_1+n_2)$ respectively.

Numerical Examples

In this section, numerical examples are presented to assess the credibility of the maximum entropy analytical approach and also to illustrate how critically the distributional form of interarrival and service times with known first two moments affects system performance in terms of primary measures such as throughput and mean queue length. The results are compared with those obtained by simulation. The simulation was based on the GE distribution representing general distributions with known first two moments. The simulation results quoted are produced using SIMSCRIPT II.5 (Russell, 1989). All measurements of time are in second. The results for the G/G/c infinite capacity queue with $c = 2, 3$ and 4 parallel servers and two classes of users are presented in Figures 48–53 respectively. Figures 54–61 represent the results for the G/G/c/N queue for $c = 2, 3$ and 4 parallel servers. SIM represents the simulation done on the model shown in figure 41 while ME represents the results obtained using the proposed maximum entropy analytical approach.

The comparative study focuses on the throughput, mean response time and mean queue length of centre 0 for the networks shown in Figure 41. In the first example, $\lambda_2 = 0.2$ (taken to be fixed for validation purposes), various values of λ_1 , C_μ^2 takes the value 0.5, $C_{\lambda_1}^2 = C_{\lambda_2}^2 = 2.0$. In the second example, $C_\mu^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = C_{\lambda_3}^2 = 0.5$

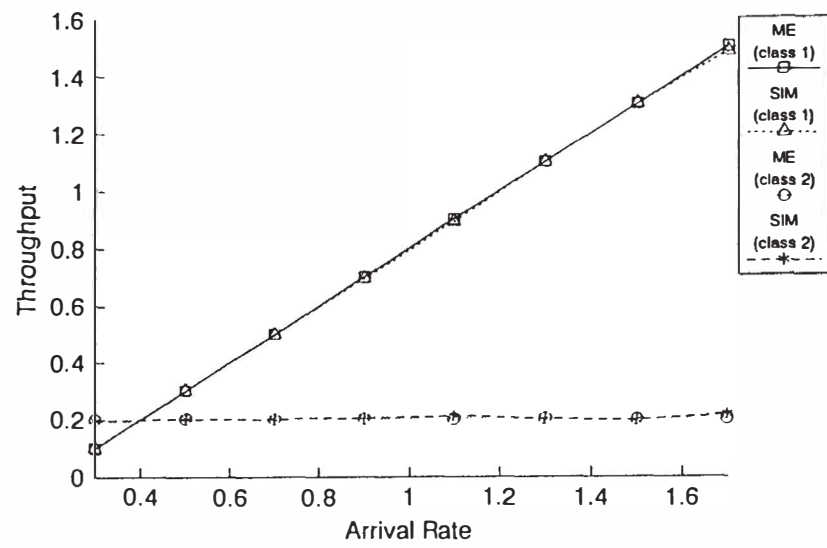


Figure 48 : Throughput of Class 1 and 2 Users versus Total Offered Load ($G/G/2$, $C_{\mu}^2 = 0.5$, $C_{\lambda_1}^2 = C_{\lambda_2}^2 = 2.0$)

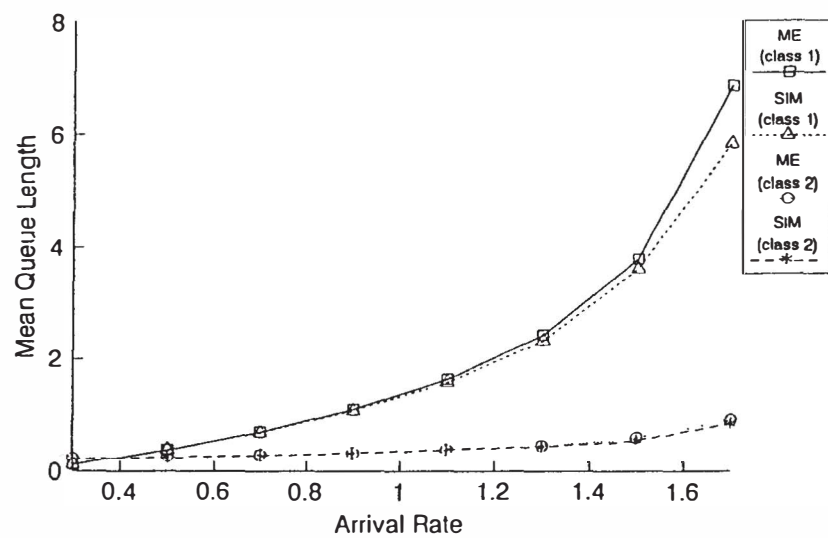


Figure 49 : Mean Queue Length of Class 1 and 2 Users versus Total Offered Load ($G/G/2$, $C_{\mu}^2 = 0.5$, $C_{\lambda_1}^2 = C_{\lambda_2}^2 = 2.0$)

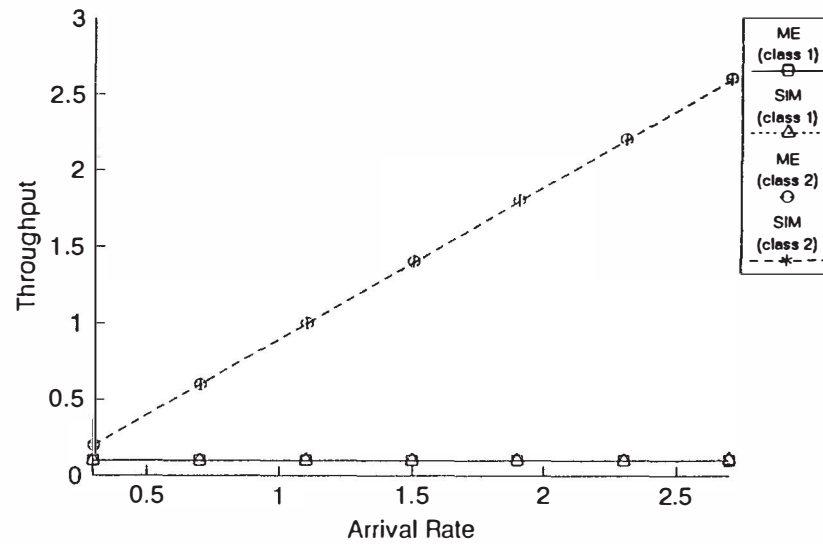


Figure 50 :Throughput of Class 1 and 2 Users versus Total Offered Load ($G/G/3$, $C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 0.5$)

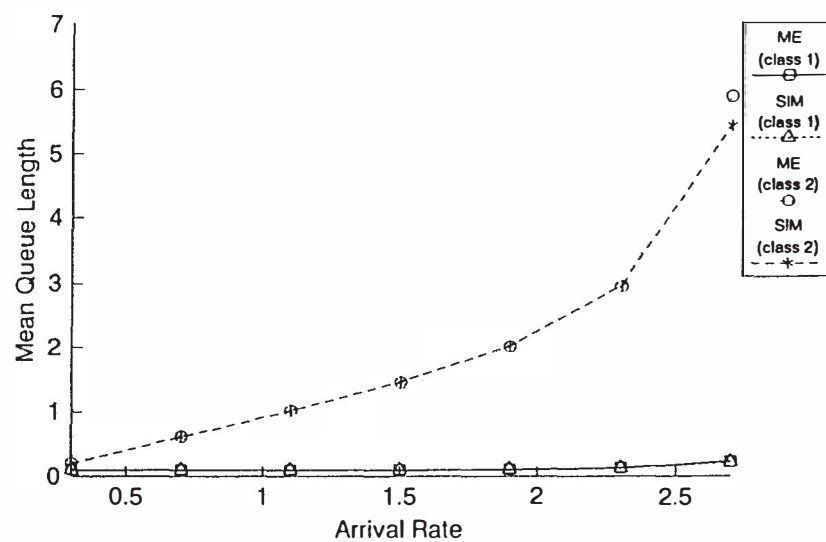


Figure 51 : Mean Queue Length of Class 1 and 2 Users versus Total Offered Load ($G/G/3$, $C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 0.5$)

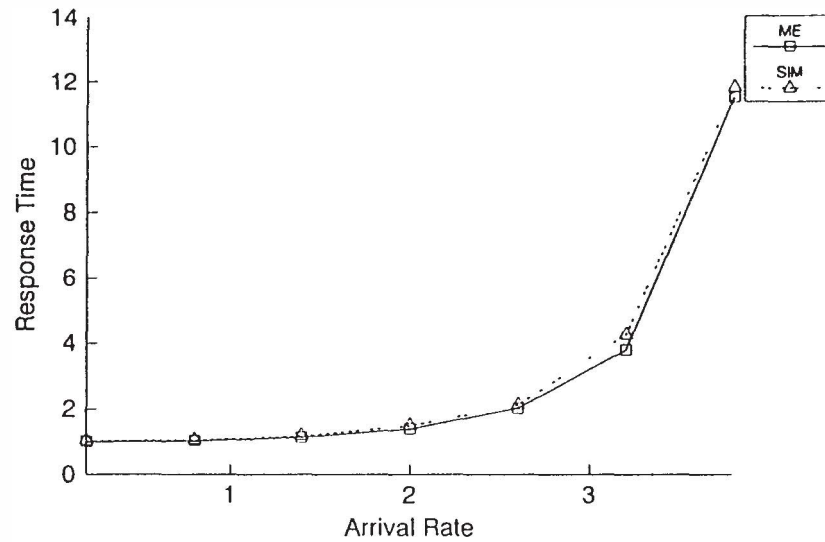


Figure 52 : Mean Response Time of Class 1 and 2 Users versus Total Offered Load ($G/G/4$, $C_{\mu}^2 = 5.0$, $C_{\lambda_1}^2 = C_{\lambda_2}^2 = 2.0$)

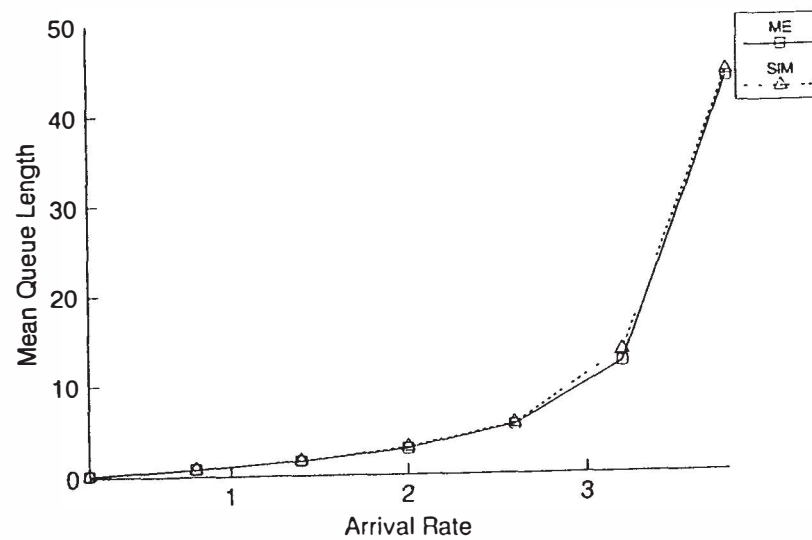


Figure 53 : Mean Queue Length of Class 1 and 2 Users versus Total Offered Load ($G/G/4$, $C_{\mu}^2 = 5.0$, $C_{\lambda_1}^2 = C_{\lambda_2}^2 = 2.0$)

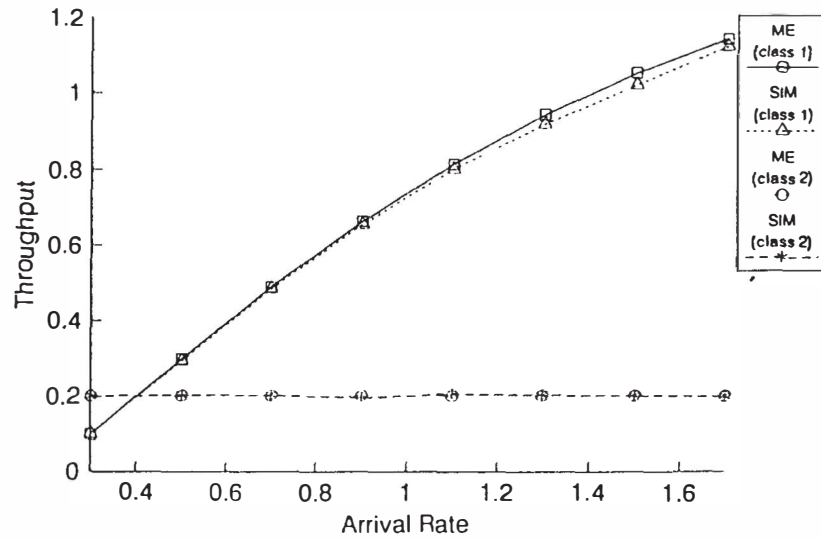


Figure 54 : Throughput of Class 1 and 2 Users versus Total Offered Load ($G/G/2/N$, $N_1 = 5$, $N_2 = 10$,

$$C_{\mu}^2 = 5.0, C_{\lambda_1}^2 = 2.0, C_{\lambda_2}^2 = 1.0)$$

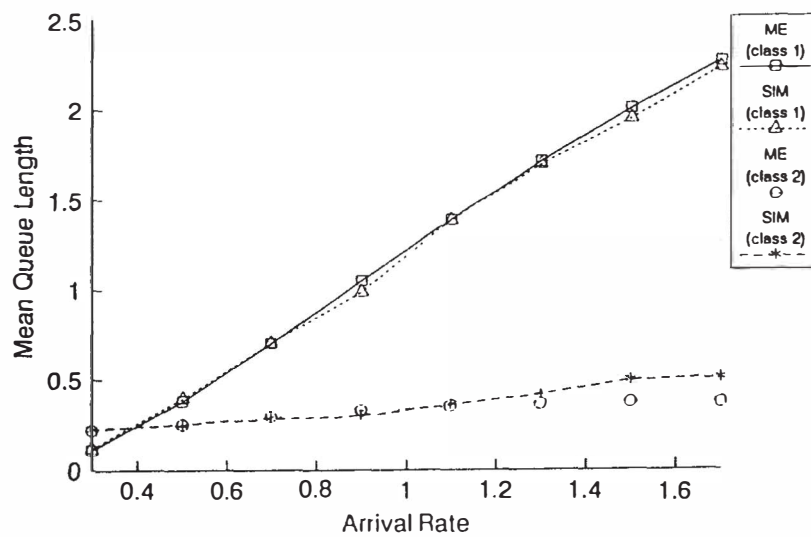


Figure 55 : Mean Queue Length of Class 1 and 2 Users versus Total Offered Load ($G/G/2/N$, $N_1 = 5$, $N_2 = 10$,

$$C_{\mu}^2 = 5.0, C_{\lambda_1}^2 = 2.0, C_{\lambda_2}^2 = 1.0)$$

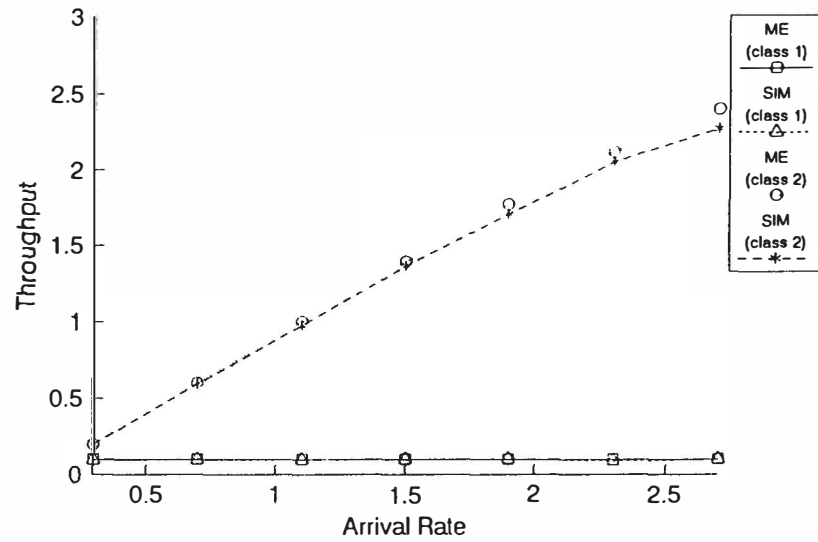


Figure 56 : Throughput of Class 1 and 2 Users versus Total Offered Load ($G/G/3/N$, $N_1 = 5$, $N_2 = 10$,

$$C_{\mu}^2 = 0.5, C_{\lambda_1}^2 = C_{\lambda_2}^2 = 2.0)$$

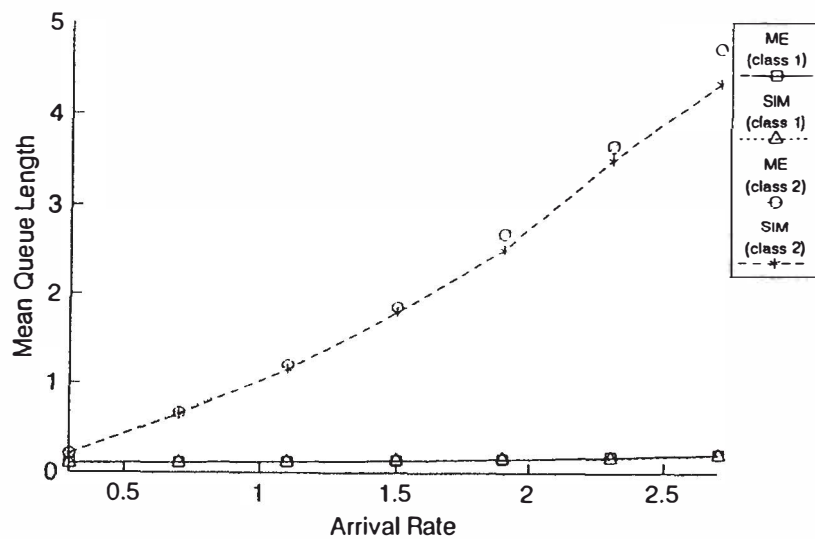


Figure 57 : Mean Queue Length of Class 1 and 2 Users versus Total Offered Load ($G/G/3/N$, $N_1 = 5$,

$$N_2 = 10, C_{\mu}^2 = 0.5, C_{\lambda_1}^2 = C_{\lambda_2}^2 = 2.0)$$

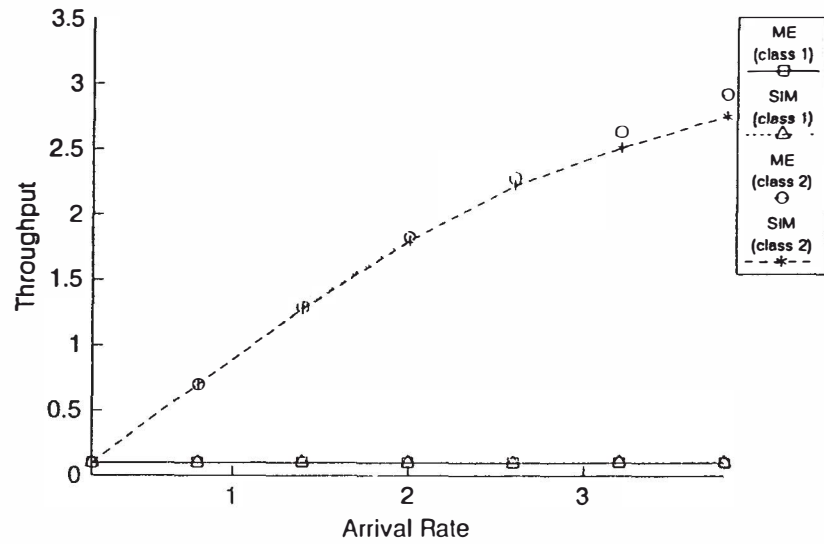


Figure 58 : Throughput of Class 1 and 2 Users versus Total Offered

Load ($G/G/4/N$, $N_1 = N_2 = 5$, $C_\mu^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 1.0$)

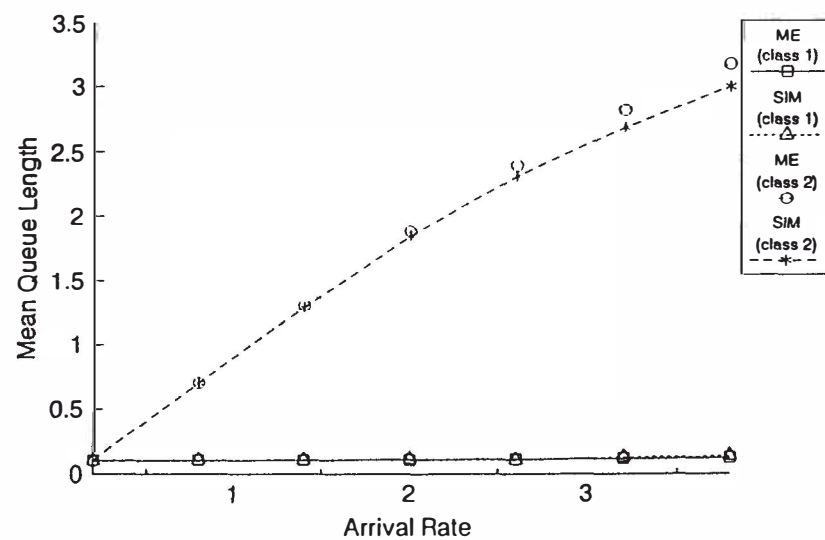


Figure 59 : Mean Queue Length of Class 1 and 2 Users versus Total

Offered Load ($G/G/4/N$, $N_1 = N_2 = 5$, $C_\mu^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 1$).

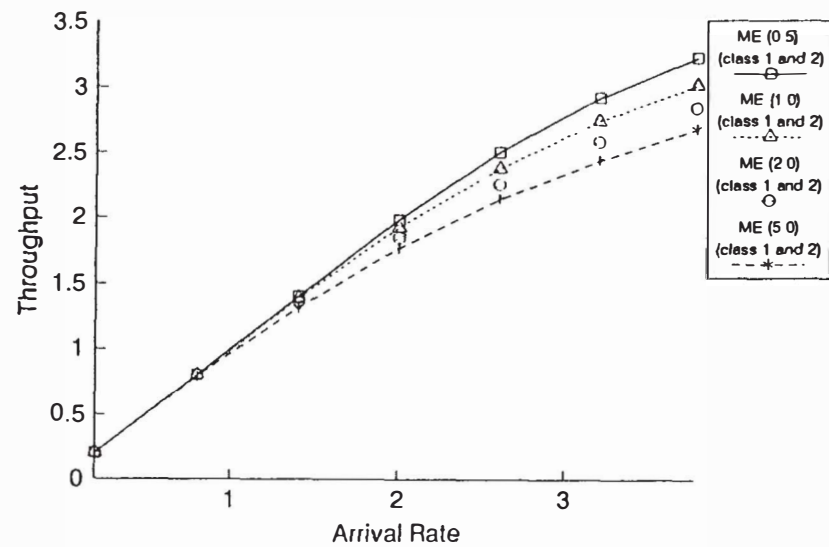


Figure 60 : Overall Throughput versus Total Offered Load
($G/G/4/N$, $N_1 = N_2 = 5$)

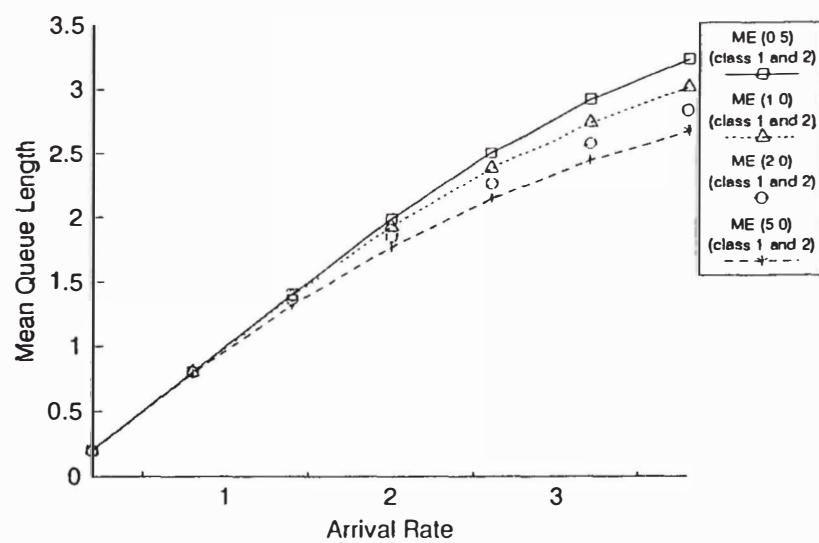


Figure 61 : Overall Mean Queue Length versus Total Offered Load
($G/G/4/N$, $N_1 = N_2 = 5$)

(see Figures 50-51) while in the third example (see Figures 52-53), $C_{\mu}^2 = 5.0$ and $C_{\lambda_1}^2 = C_{\lambda_2}^2 = 2.0$. In both the second and third example, $\lambda_1 = 0.1$ for various values of λ_2 . In the fourth example, $C_{\mu}^2 = 5.0$, $C_{\lambda_1}^2 = 2.0$, $C_{\lambda_2}^2 = 1.0$ and $\lambda_2 = 0.2$ for various values of λ_1 (see Figures 54-55) while in the fifth example, $C_{\mu}^2 = 0.5$, $C_{\lambda_1}^2 = C_{\lambda_2}^2 = 2.0$ and $\lambda_1 = 0.2$ for various values of λ_2 (see Figures 56-57). In both the fourth and fifth examples, $N_1 = 5$ and $N_2 = 10$. In the sixth example, $N_1 = N_2 = 5$, $\lambda_1 = 0.1$ for various values of λ_2 and $C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 1.0$ (see Figures 58-59). In all the above examples the throughput, response time and mean queue length are plotted against various offered load. It can be generally observed that the ME results are comparable to those obtained by SIM under all traffic condition.

It is also noted that the GE type solution (i.e. in example six) for the overall throughput as a function of the total offered load provides lower bounds (pessimistic) when $C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 5.0$ and upper bounds (optimistic) when $C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 0.5$ (see Figure 60). As expected, the overall mean queue length as a function of the total offered load provide lower bounds (pessimistic) when $C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 0.5$ and upper bounds (optimistic) when $C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 5.0$ (see Figure 61).

Summary

The principle of maximum entropy has been used to derive a new product form solution for the approximate analysis of a stable

multiclass $G/G/c$ queue with general class interarrival and service time distributions and $c \geq 2$ parallel servers under FCFS service discipline. The generalised exponential (GE) distributional model is used to represent general distributions with known first two moments per class. Numerical examples illustrate the accuracy of the proposed approximations in relation to simulations involving different interarrival and service time distributions per class. Useful insights into how critically the system performance is affected by the distributional form of the service times are also included.

CHAPTER VI
DECENTRALISED OPTIMAL FLOW CONTROL OF SINGLE SERVER
GENERAL QUEUEING NETWORKS WITH MULTICLASS USERS

Introduction

This chapter is concerned with the decentralised optimal flow control of computer communication networks with general packet interarrival and transmission time distributions. Flow control is intended to provide fast and reliable communication with efficient allocation of network resources. The performance of a network can be greatly degraded by high utilisation links (bottlenecks). One purpose of flow control is to make efficient use of network resources by limiting or rerouting traffic, thus avoiding bottlenecks.

Recent work on optimal flow control in computer communication network has been focused on the centralised control of a single class of users. The centralised optimal flow control for a class of queueing networks at equilibrium is of the window type (Lazar, 1982a; 1983b; Kouvatsos and Othman, 1989c; 1989d). The authors used an optimality criterion, based on an earlier suggestion by Schwartz (1980), which for an offered load maximizes the throughput under the constraint of the average time delay not exceeding a preassigned value. This kind of analysis is necessary to incorporate flow control policies in communication networks which take into consideration not only the offered load but also the maximum time delay. Although small networks can be controlled in such a manner, the centralised approach is inefficient due to the large overhead required for obtaining the necessary information at the central controller. It is clearly

desirable therefore, to implement flow control procedures that are decentralised in nature. In this environment, users are able to execute flow control procedures based on local information only.

Hsiao and Lazar (1984) studied the decentralised local flow control in computer communication networks with exponential packet interarrival and transmission patterns, and two classes of users sharing a high utilisation link. The decentralised flow control of the link is shown to be a pair of window type mechanisms where the optimal window sizes depend on the maximum packet generation rate of the two users, the average service rate and the maximum tolerable average time delay. However, it is also essential to extend these results in order to investigate how critically the flow control strategy and the optimal throughput are affected by the distributional form of packet interarrival and transmission patterns in communication networks.

Using Hsiao and Lazar (1984) optimality criterion, this chapter examines the decentralised optimal flow control of a high utilisation link in computer communication networks with general packet interarrival and transmission pattern. Since one of the main design requirements for communication protocols is the acknowledgment of packets reaching their destination, a general closed queueing network is used as a model for studying the decentralised optimal flow control problem (Lam and Lien, 1981).

The decentralised flow control problem considered here assumes that resources are completely shared. User packets queue up in a common buffer to wait for transmission. The decentralised optimal flow control is obtained for a larger class of control mechanisms, which incorporate local information on the number of outstanding packets.

The optimality criterion also allows explicit comparison of the throughput time-delay trade-off. In view of the multiplicity of user classes, there exist naturally more than one objective to be considered (Lazar, 1982b). This chapter considers the global objective of maximizing the average throughput subject to a bound on the average time delay over all the users sharing the link.

This chapter is organised as follows. The model for the bottleneck decentralised flow control is presented and the maximum entropy GE type solution for the approximation of the queue length distribution of a G/G/1/N with two classes of users is introduced. Relevant optimisation problem is then defined using a global objective function. Numerical examples will then be provided followed by a summary of this chapter.

The Mathematical Model

The model for the decentralised control problem is shown in Figure 62. In the following, the bottleneck is modeled as a general server with fixed service rate μ , constant coefficient of variation C_μ and a FCFS service discipline. K classes of users are considered. Packets belonging to either class queue up in a common buffer to wait for transmission via the bottleneck link. User class i , $i = 1, 2, \dots, K$, packets are generated at a maximum rate c_i packets per second. The decentralised optimal flow control problem is formulated as follows. Each user controls his rate of packet generation using information of the number of outstanding packets (i.e via acknowledgment or otherwise). Information on the state of the other user class is not available. It is assumed that the number of class i packets is N_i

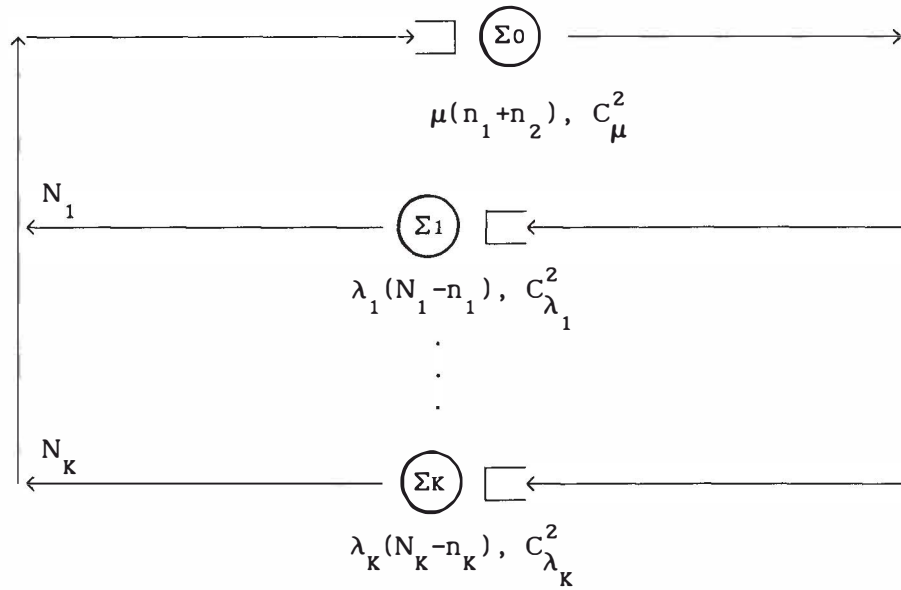


Figure 62 : Model for Bottleneck Decentralised Flow Control

where N_i are arbitrarily large integers and the total number of packets in the network is $N = N_1 + N_2 + \dots + N_K$.

The model in Figure 62 is a general closed queueing network which has been studied in Chapter IV. Here, $\underline{n} = (n_1, n_2, \dots, n_K)$ is the state of the server (i.e link), n_i being the number class i packets in the queue, respectively. $\lambda_i(N_i - n_i)$ is the corresponding controlled packet generation rates with n_i outstanding packets.

The mean throughput and the mean time delay of the bottleneck link of Figure 62 are given by

$$\langle \chi \rangle_N = \sum_{n_1=0}^{N_1} \dots \sum_{n_K=0}^{N_K} \mu(n_1+n_2+\dots+n_K) p_N(\underline{n}) \quad [6.1]$$

$(n_1+\dots+n_K > 0)$

and

$$\langle \tau \rangle_N = \frac{\sum_{n_1=0}^{N_1} \dots \sum_{n_K=0}^{N_K} (n_1 + n_2 + \dots + n_K) p_N(\underline{n})}{\sum_{n_1=0}^{N_1} \dots \sum_{n_K=0}^{N_K} \mu(n_1 + n_2 + \dots + n_K) p_N(\underline{n})}, \quad [6.2]$$

(n₁ + ... + n_K > 0)

respectively, where $p_N(\underline{n})$, is the corresponding joint queue length distribution of the bottleneck link of Figure 62 and can be analogously expressed by

$$p_N(\underline{n}) = p_N(0) \binom{n_1 + n_2 + \dots + n_K}{n_1} \left[\prod_{j=0}^{n_1-1} \frac{a_1(j, n_2, n_3, \dots, n_K)}{b_1(j+1, n_2, n_3, \dots, n_K)} \right] \binom{n_2 + n_3 + \dots + n_K}{n_2} \left[\prod_{j=0}^{n_2-1} \frac{a_2(0, j, n_3, \dots, n_K)}{b_2(0, j+1, n_3, \dots, n_K)} \right] \dots \left[\prod_{j=0}^{n_K-1} \frac{a_K(0, 0, \dots, 0, j)}{b_K(0, 0, \dots, 0, j+1)} \right], \quad [6.3]$$

$$\begin{aligned} n_i &= 1, 2, \dots, N_i, \\ i &= 1, 2, \dots, K, \\ n_1 + n_2 + \dots + n_K &\neq 0, \end{aligned}$$

and

$$p_N(0) = \left[1 + \sum_{n_1=0}^{N_1} \dots \sum_{n_K=0}^{N_K} p_N(\underline{n}) \right]^{-1}, \quad [6.4]$$

(n₁ + ... + n_K > 0)

where $a_i(\underline{n})$ and $b_i(\underline{n})$ ($i = 1, 2, \dots, K$), are the 'up' and 'down' rates. Due to the complexity involved in writing down $a_i(\underline{n})$ and $b_i(\underline{n})$ for K classes of users, we will assume that only two classes (i.e. $K = 2$) of users share the service of the link on a FCFS discipline. The

solutions can easily be extended to more than two classes of users. The joint queue length distribution of the bottleneck link of Figure 62 for two classes of users can be expressed by

$$p_N(n_1, n_2) = \binom{n_1+n_2}{n_1} \left[\prod_{j=0}^{n_1-1} \frac{a_1(j, n_2)}{b_1(j+1, n_2)} \right] \left[\prod_{j=0}^{n_2-1} \frac{a_2(0, j)}{b_2(0, j+1)} \right] p_N(0, 0), \quad [6.5]$$

and

$$p_N(0, 0) = \left[1 + \sum_{\substack{n_1=0 \\ (n_1+n_2>0)}}^{N_1} \sum_{n_2=0}^{N_2} \binom{n_1+n_2}{n_1} \left[\prod_{j=0}^{n_1-1} \frac{a_1(j, n_2)}{b_1(j+1, n_2)} \right] \left[\prod_{j=0}^{n_2-1} \frac{a_2(0, j)}{b_2(0, j+1)} \right] \right]^{-1}, \quad [6.6]$$

where $\prod_{j=0}^{-1} = 1$ and the parameters $a_1(n_1, n_2)$, $b_1(n_1, n_2)$, $a_2(0, n_2)$ and $b_2(0, n_2)$ represent 'up' and 'down' flow coefficients (see Figure 63) given by

$$a_1(n_1, n_2) = \begin{cases} \pi \lambda(N) \tau_1 \tau_2, & n_1 = 0, n_2 = 0, \\ \pi e, & \left. \begin{array}{l} n_1 = 0, 1, \dots, N_1-1, \\ n_2 = 0, 1, \dots, N_2, \end{array} \right\} n_1+n_2>0, \end{cases} \quad [6.7]$$

$$b_1(n_1, n_2) = \begin{cases} f, & \begin{array}{l} n_1=1, 2, \dots, N_1-1, \\ n_2=0, 1, \dots, N_2, \end{array} \\ \frac{f \tau_1 \tau_2 [\mu(n_1+n_2) - [1-\pi] \lambda(N-n_1-n_2)]}{f - [1-\pi] e}, & \begin{array}{l} n_1=N_1, \\ n_2=0, 1, \dots, N_2-1, \end{array} \\ \mu(n_1+n_2) \tau_2 - \pi [\mu(n_1+n_2) \tau_2 [1-\tau_1]], & n_1=N_1, n_2=N_2, \end{cases} \quad [6.8]$$

$$a_2(0, n_2) = \begin{cases} [1-\pi] \lambda(N) \tau_1 \tau_2, & n_2 = 0, \\ [1-\pi] e, & n_2 = 0, 1, \dots, N_2-1, \end{cases} \quad [6.9]$$

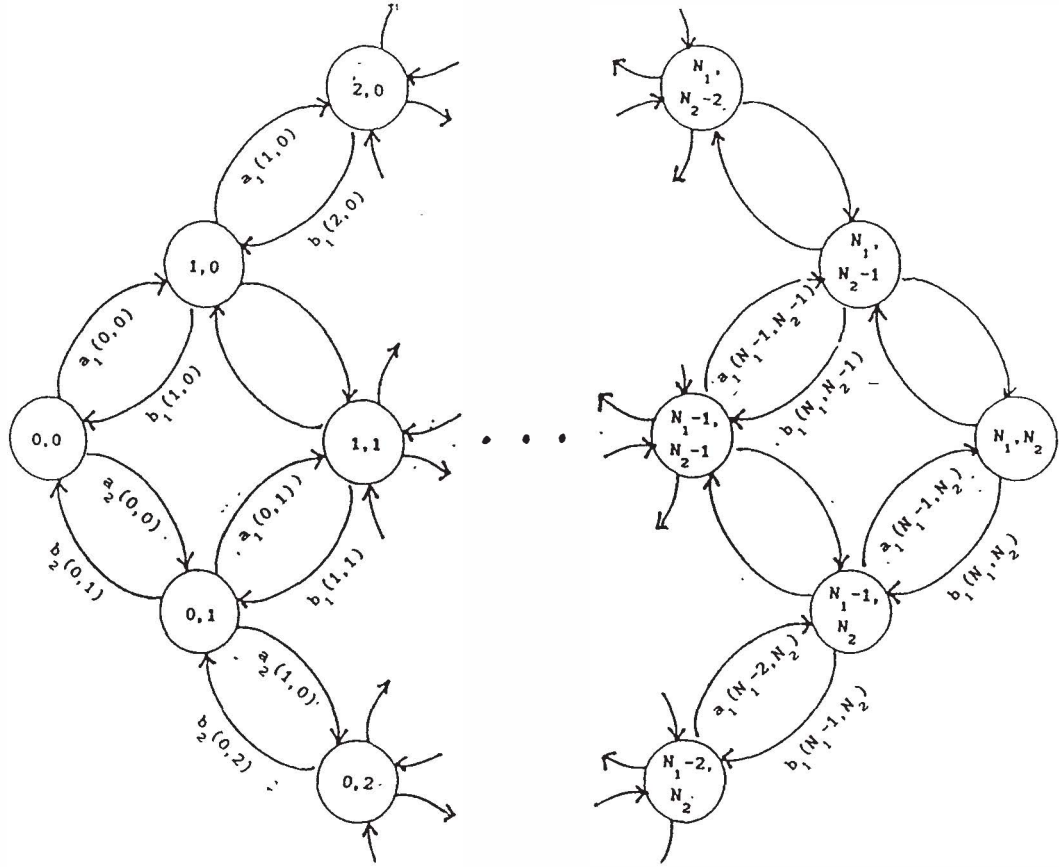


Figure 63 : Transition 'Up' and 'Down' Rate of
GE(n)/GE(n)/1/N with Two Classes of Users

$$b_2(n_1, n_2) = \begin{cases} f, & n_2 = 0, 1, \dots, N_2-1, \\ \frac{f \tau_1 \tau_2 [\mu(n_1 + n_2) - \pi \lambda (N - n_1 - n_2)]}{f - \pi e}, & n_2 = N_2, \end{cases} \quad [6.10]$$

where

$$e = \mu(n_1 + n_2) \tau_2 [1 - \tau_1] + \lambda(N - n_1 - n_2) \tau_1,$$

$$f = \mu(n_1 + n_2) \tau_2 + \tau_1 [1 - \tau_2] \lambda(N - n_1 - n_2).$$

$$\pi = \frac{\lambda_1 (N_1 - n_1)}{\lambda(N - n_1 - n_2)}, \quad \tau_1 = \frac{2}{(C_a^2 + 1)}, \quad \tau_2 = \frac{2}{(C_\mu^2 + 1)},$$

and

$$C_a^2 = \frac{\lambda(N-n_1-n_2)}{\frac{\lambda_1(N_1-n_1)}{C_{\lambda_1}^2 + 1} + \frac{\lambda_2(N_2-n_2)}{C_{\lambda_2}^2 + 1}} - 1.$$

Note that the joint queue length distribution of the GE(n)/GE(n)/1/N queue can be interpreted as the joint steady state probability of station 0 containing n (i.e. $n = n_1 + n_2$) packets. The versatile form of the joint queue length distribution is used in the next section to formulate the optimality criterion. For simplicity we focus on class 1 user packets and the results for class 2 user packets are symmetrical to those obtained for class 1 user packets.

The Optimisation Criterion

Consider the control parameters of the general cyclic queue of Figure 62. Let us denote $\lambda_1 = \{\lambda_1(n_1)\}$ for $n_1 = 1, 2, \dots, N_1$, and $\lambda_2 = \{\lambda_2(n_2)\}$ for $n_2 = 1, 2, \dots, N_2$, as a set of control parameters which satisfy the peak constraints

$$0 \leq \lambda_1(n_1) \leq c_1 \quad [6.11]$$

$$0 \leq \lambda_2(n_2) \leq c_2 \quad [6.12]$$

for all n_1 and n_2 , for $n_1 = 1, 2, \dots, N_1$, $n_2 = 1, 2, \dots, N_2$, where c_1 and $c_2 \in \mathbb{R}_+$ are constant called admissible. The maximum input rate rates c_1 and c_2 offered by the source nodes are assumed to be known.

The set of control parameters $\lambda_1 = \{\lambda_1(n_1)\}$ for $n_1 = 1, 2, \dots, N_1$, and $\lambda_2 = \{\lambda_2(n_2)\}$ for $n_2 = 1, 2, \dots, N_2$, is said to be optimum over the class of admissible controls for a given time delay $T \in \mathbb{R}_+$, if it achieves the maximum

$$F(T) = \max \langle \chi \rangle_N \quad [6.13]$$

subject to

$$\langle \tau \rangle_N \leq T \quad [6.14]$$

Here $F(T)$ is called the throughput time delay function. For the sake of simplicity, in what follows only homogeneous service times will be considered (i.e. $\mu(n_1+n_2) = \mu$ for all n_1 and n_2 ; $0 \leq n_1 \leq N_1$, $0 \leq n_2 \leq N_2$). Let us denote the expressions

$$x_1(n_1, n_2) = \prod_{j=0}^{n_1-1} \frac{a_1(j, n_2)}{b_1(j+1, n_2)}, \quad [6.15]$$

and

$$x_2(0, n_2) = \prod_{j=0}^{n_2-1} \frac{a_2(0, j)}{b_2(0, j+1)}, \quad [6.16]$$

or can be written in a step recursion as

$$x_1(n_1, n_2) = \frac{a_1(n_1-1, n_2)}{b_1(n_1, n_2)} x_1(n_1-1, n_2), \quad \begin{array}{l} n_1 = 1, 2, \dots, N_1, \\ n_2 = 0, 1, \dots, N_2, \end{array} \quad [6.17]$$

and

$$x_2(0, n_2) = \frac{a_2(0, n_2 - 1)}{b_2(0, n_2)} x_2(0, n_2), \quad n_2 = 1, 2, \dots, N_2, \quad [6.18]$$

where $x_1(0, n_2) = x_2(0, 0) = 1$, $n_2 = 0, 1, \dots, N_2$. Hence for a GE(n)/GE/1/N queue we have

$$x_1(n_1, n_2) = \left\{ \begin{array}{ll} \pi \left[\frac{\lambda(N) \tau_1 \tau_2}{\mu \tau_2 + \lambda(N-1) \tau_1 [1-\tau_2]} \right], & n_1 = 1, n_2 = 0, \\ x_1(n_1 - 1, n_2) \pi \left[\frac{\mu \tau_2 [1-\tau_1] + \lambda(N-n_1-n_2+1) \tau_1}{\mu \tau_2 + \lambda(N-n_1-n_2) \tau_1 [1-\tau_2]} \right], & \left. \begin{array}{l} n_1 = 1, 2, \dots, N_1 - 1, \\ n_2 = 0, 1, \dots, N_2, \end{array} \right\} n_1 + n_2 > 1 \\ x_1(n_1 - 1, n_2) \pi \left[\frac{\mu \tau_2 [1-\tau_1] + \lambda(N-n_1-n_2+1) \tau_1}{\mu \tau_2 + \lambda(N-n_1-n_2) \tau_1 [1-\tau_2]} \right] & \\ \frac{\mu \tau_2 + \lambda(N-n_1-n_2) \tau_1 [1-\tau_2] - [1-\pi] \left[\mu \tau_2 [1-\tau_1] + \lambda(N-n_1-n_2) \tau_1 \right]}{\tau_1 \tau_2 \left[\mu - [1-\pi] \lambda(N-n_1-n_2) \right]}, & \left. \begin{array}{l} n_1 = N_1, \\ n_2 = 0, 1, \dots, N_2 - 1, \end{array} \right\} \\ x_1(N_1 - 1, N_2) \pi \left[\frac{\mu \tau_2 [1-\tau_1] + \lambda(1) \tau_1}{\mu \tau_2 - \pi \mu \tau_2 [1-\tau_1]} \right] & \left. \begin{array}{l} n_1 = N_1, \\ n_2 = N_2, \end{array} \right\} \end{array} \right. \quad [6.19]$$

and

$$\begin{aligned}
& x_2(0, n_2) = \\
& \left\{ \begin{aligned}
& [1-\pi] \left[\frac{\lambda(N)\tau_1\tau_2}{\mu\tau_2 + \lambda(N-1)\tau_1[1-\tau_2]} \right], \quad n_2 = 1, \\
& x_2(0, n_2-1) [1-\pi] \left[\frac{\mu\tau_2[1-\tau_1] + \lambda(N-n_2+1)\tau_1}{\mu\tau_2 + \lambda(N-n_2)\tau_1[1-\tau_2]} \right], \\
& \quad n_2 = 2, 3, \dots, N_2-1, \\
& x_2(0, n_2-1) [1-\pi] \left[\frac{\mu\tau_2[1-\tau_1] + \lambda(N-n_2+1)\tau_1}{\mu\tau_2 + \lambda(N-n_2)\tau_1[1-\tau_2]} \right] \\
& \quad \frac{\mu\tau_2 + \lambda(N-n_2)\tau_1[1-\tau_2] - \pi[\mu\tau_2[1-\tau_1] + \lambda(N-n_2)\tau_1]}{\tau_1\tau_2[\mu - \pi\lambda(N-n_2)]}, \quad n_2 = N_2.
\end{aligned} \right. \quad [6.20]
\end{aligned}$$

Thus it can be verified that [6.5] and [6.6] can be written as

$$p_N(n_1, n_2) = \binom{n_1+n_2}{n_1} x_1(n_1, n_2) x_2(0, n_2) p_N(0, 0), \quad [6.21]$$

and

$$p_N(0, 0) = \left[1 + \sum_{n_1=0}^{N_1} \sum_{\substack{n_2=0 \\ (n_1+n_2>0)}}^{N_2} \binom{n_1+n_2}{n_1} x_1(n_1, n_2) x_2(0, n_2) \right]^{-1}, \quad [6.22]$$

respectively. Thus it follows that

$$\begin{aligned}
& \sum_{n_1=0}^{N_1} \sum_{\substack{n_2=0 \\ (n_1+n_2>0)}}^{N_2} \binom{n_1+n_2}{n_1} x_1(n_1, n_2) x_2(0, n_2) = \frac{\sum_{n_1=0}^{N_1} \sum_{\substack{n_2=0 \\ (n_1+n_2>0)}}^{N_2} p_N(n_1, n_2)}{p_N(0, 0)} = \frac{1}{p_N(0, 0)} - 1.
\end{aligned} \quad [6.23]$$

It can easily be seen that via [6.23]

$$\max \langle \chi \rangle_N = \max_{\mu} \sum_{n_1=0}^{N_1} \sum_{\substack{n_2=0 \\ (n_1+n_2>0)}}^{N_2} p_N(n_1, n_2) = \max_{\mu} \mu [1 - p_N(0,0)] \quad [6.24]$$

subject to $\langle \tau \rangle_N \leq T$, i.e. maximising the throughput by using a bounded time delay criterion is equivalent to minimising the probability that the system is idle. Thus the optimisation criterion can be described by means of the following lemma.

Lemma 6.1

The control parameters $\lambda_1 = \lambda_1(n_1)$, $\lambda_2 = \lambda_2(n_2)$, $n_1 = 1, 2, \dots, N_1$, $n_2 = 1, 2, \dots, N_2$, are optimum in the set of admissible controls for a given time delay T if they achieved the maximum of

$$\max \left\{ \sum_{n_1=0}^{N_1} \sum_{\substack{n_2=0 \\ (n_1+n_2>0)}}^{N_2} \binom{n_1+n_2}{n_1} x_1(n_1, n_2) x_2(0, n_2) = \frac{1}{p_N(0,0)} - 1 \right\} \quad [6.25]$$

subject to

$$\sum_{n_1=0}^{N_1} \sum_{\substack{n_2=0 \\ (n_1+n_2>0)}}^{N_2} (n_1+n_2 - \mu T) \binom{n_1+n_2}{n_1} x_1(n_1, n_2) x_2(0, n_2) \leq 0 \quad [6.26]$$

where

$$x_1(n_1, n_2) \leq \hat{x}_1(n_1, n_2) \quad [6.27]$$

and

$$x_2(0, n_2) \leq \hat{x}_2(0, n_2) \quad [6.28]$$

for $n_1 + n_2 = 1, 2, \dots, N$ and

$$\hat{x}_1(n_1, n_2) = \begin{cases} \pi \left[\frac{\lambda_{\max} \tau_1 \tau_2}{\mu \tau_2 + \lambda_{\max} \tau_1 [1 - \tau_2]} \right], & n_1 = 1, n_2 = 0, \\ \hat{x}_1(n_1 - 1, n_2) \pi \left[\frac{\mu \tau_2 [1 - \tau_1] + \lambda_{\max} \tau_1}{\mu \tau_2 + \lambda_{\max} \tau_1 [1 - \tau_2]} \right], & n_1 = 1, 2, \dots, N_1 - 1, \\ & n_2 = 0, 1, \dots, N_2, \\ \hat{x}_1(n_1 - 1, n_2) \pi \left[\frac{\mu \tau_2 [1 - \tau_1] + \lambda_{\max} \tau_1}{\mu \tau_2 + \lambda_{\max} \tau_1 [1 - \tau_2]} \right] \\ \frac{\mu \tau_2 + \lambda_{\max} \tau_1 [1 - \tau_2] - [1 - \pi] \left[\mu \tau_2 [1 - \tau_1] + \lambda_{\max} \tau_1 \right]}{\tau_1 \tau_2 [\mu - [1 - \pi] \lambda_{\max}]}, & n_1 = N_1, \\ & n_2 = 0, 1, \dots, N_2 - 1, \\ \hat{x}_1(N_1 - 1, N_2) \pi \left[\frac{\mu \tau_2 [1 - \tau_1] + \lambda_{\max} \tau_1}{\mu \tau_2 - \pi \mu \tau_2 [1 - \tau_1]} \right], & n_1 = N_1, \\ & n_2 = N_2, \end{cases} \quad [6.29]$$

and

$$\hat{x}_2(0, n_2) = \begin{cases} [1-\pi] \left[\frac{\lambda_{\max} \tau_1 \tau_2}{\mu \tau_2 + \lambda_{\max} \tau_1 [1-\tau_2]} \right], & n_2 = 1, \\ \hat{x}_2(0, n_2 - 1) [1-\pi] \left[\frac{\mu \tau_2 [1-\tau_1] + \lambda_{\max} \tau_1}{\mu \tau_2 + \lambda_{\max} \tau_1 (1-\tau_2)} \right], & n_2 = 2, 3, \dots, N_2 - 1, \\ \hat{x}_2(0, n_2 - 1) [1-\pi] \left[\frac{\mu \tau_2 [1-\tau_1] + \lambda_{\max} \tau_1}{\mu \tau_2 + \lambda_{\max} \tau_1 [1-\tau_2]} \right] \\ \frac{\mu \tau_2 + \lambda_{\max} \tau_1 [1-\tau_2] - \pi \left[\mu \tau_2 [1-\tau_1] + \lambda_{\max} \tau_1 \right]}{\tau_1 \tau_2 [\mu - \pi \lambda_{\max}]}, & n_2 = N_2. \end{cases} \quad [6.30]$$

Proof

From [6.14] and [6.21] we have

$$\begin{aligned} \langle \tau \rangle_N &= \frac{\sum_{n_1=0}^{N_1} \sum_{n_2=0}^{N_2} (n_1 + n_2) p_N(n_1, n_2)}{\sum_{n_1=0}^{N_1} \sum_{n_2=0}^{N_2} \mu p_N(n_1, n_2)} = \\ &= \frac{\sum_{n_1=0}^{N_1} \sum_{n_2=0}^{N_2} \binom{n_1 + n_2}{n_1} (n_1 + n_2) x_1(n_1, n_2) x_2(0, n_2)}{\mu \sum_{n_1=0}^{N_1} \sum_{n_2=0}^{N_2} \binom{n_1 + n_2}{n_1} x_1(n_1, n_2) x_2(0, n_2)} \leq T \quad [6.31] \end{aligned}$$

or equivalently, inequality [6.26] holds and from [6.1] and [6.23], it follows that the relations [6.24] and subsequently [6.25] are satisfied. Furthermore, from [6.11], [6.12], [6.19], [6.20], [6.29] and [6.31], relations [6.27] and [6.28] follow.

The Optimal Control

Some results will first be presented in order to facilitate the proof of the main theorem on the optimal control that follows. In Lemma 6.2 we derive the expression for the maximum time delay and an achievable upper bound on it. In Lemma 6.3 it is shown that the maximum achievable time delay in the GE(n)/GE/1/N system is increasing with the number of packets.

Lemma 6.2

The maximum mean time delay $T_{\max}(L_1, L_2)$ achieved with L_1 and L_2 of class 1 and 2 packets in the receiving station is given by

$$T_{\max}(L_1, L_2) = \frac{\sum_{n_1=0}^{L_1} \sum_{n_2=0}^{L_2} \binom{n_1+n_2}{n_1} [n_1+n_2] \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)}{\mu \sum_{\substack{n_1=0 \\ (n_1+n_2>0)}}^{L_1} \sum_{n_2=0}^{L_2} \binom{n_1+n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)} \leq \frac{L_1 + L_2}{\mu}, \quad [6.32]$$

Proof

Let us first show that

$$\begin{aligned}
& \frac{\sum_{n_1=0}^{L_1} \sum_{n_2=0}^{L_2} \binom{n_1+n_2}{n_1} (n_1+n_2) x_1(n_1, n_2) x_2(0, n_2)}{\sum_{n_1=0}^{L_1} \sum_{n_2=0}^{L_2} \binom{n_1+n_2}{n_1} x_1(n_1, n_2) x_2(0, n_2)} \leq \\
& \frac{\sum_{n_1=0}^{L_1} \sum_{n_2=0}^{L_2} \binom{n_1+n_2}{n_1} (n_1+n_2) \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)}{\sum_{n_1=0}^{L_1} \sum_{n_2=0}^{L_2} \binom{n_1+n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)} , \quad [6.33]
\end{aligned}$$

This can be written as

$$\begin{aligned}
& \frac{\sum_{n_1=0}^{L_1} \sum_{n_2=0}^{L_2} \binom{n_1+n_2}{n_1} \binom{m_1+m_2}{m_1} \left[n_1+n_2 - m_1-m_2 \right]}{\sum_{n_1=0}^{L_1} \sum_{n_2=0}^{L_2} \binom{n_1+n_2}{n_1} \binom{m_1+m_2}{m_1}} \leq 0, \quad [6.34]
\end{aligned}$$

Due to the symmetric in n_1, n_2, m_1 and m_2 and the fact that

$$\frac{x_1(n_1, n_2) x_2(0, n_2)}{x_1(m_1, m_2) x_2(0, m_2)} \leq \frac{\hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)}{\hat{x}_1(m_1, m_2) \hat{x}_2(0, m_2)} \quad [6.35]$$

or

$$\begin{aligned}
& x_1(n_1, n_2) x_2(0, n_2) \hat{x}_1(m_1, m_2) \hat{x}_2(0, m_2) \leq \\
& \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2) x_1(m_1, m_2) x_2(0, m_2) \quad [6.36]
\end{aligned}$$

for all $n_1 + n_2 > m_1 + m_2$, the left hand side of expression [6.34] is negative. Note also that

$$T_{\max}(L_1, L_2) \leq \frac{L_1 + L_2}{\mu}, \quad [6.37]$$

since inequality [6.32] can be rewritten as

$$\sum_{n_1=0}^{L_1} \sum_{n_2=0}^{L_2} \left[(L_1 + L_2) - (n_1 + n_2) \right] \binom{n_1 + n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2) \geq 0. \quad [6.38]$$

($n_1 + n_2 > 0$)

The latter inequality is true since $n_1 + n_2 \leq L_1 + L_2$. Note that the bound of $T_{\max}(L_1, L_2)$ applies to the case when the packets after receiving service instantaneously rejoin the queue. Then $(L_1 + L_2)/\mu$ is exactly the mean time delay of a GE(n)/GE/1/N queue constantly having $L_1 + L_2$ packets in the system. Moreover, the following lemma holds.

Lemma 6.3

The maximum mean time delay $T_{\max}(L_1, L_2)$ in the GE(n)/GE/1/N system is strictly increasing with the number of packets L_1 and L_2 i.e.

$$T_{\max}(L_1, L_2) < T_{\max}(L_1 + 1, L_2) \quad [6.39]$$

where $L_1 \leq N_1$ and $L_2 \leq N_2$.

Proof

It is sufficient to show that

$$\begin{aligned}
& \frac{\sum_{n_1=0}^{L_1} \sum_{n_2=0}^{L_2} \binom{n_1+n_2}{n_1} (n_1+n_2) \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)}{\sum_{n_1=0}^{L_1} \sum_{n_2=0}^{L_2} \binom{n_1+n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)} < \\
& \frac{\sum_{n_1=0}^{L_1+1} \sum_{n_2=0}^{L_2} \binom{n_1+n_2}{n_1} (n_1+n_2) \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)}{\sum_{n_1=0}^{L_1+1} \sum_{n_2=0}^{L_2} \binom{n_1+n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)}, \quad [6.40]
\end{aligned}$$

which can be written

$$\begin{aligned}
& \left[\sum_{n_1=0}^{L_1} \sum_{n_2=0}^{L_2} \binom{n_1+n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2) \right] \\
& \left[\sum_{i=0}^{L_2} \binom{L_1+1+i}{L_1+1} \left[n_1+n_2-(L_1+1+i) \right] \hat{x}_1(L_1+1, i) \hat{x}_2(0, i) \right] < 0. \quad [6.41]
\end{aligned}$$

This inequality is true since $(n_1+n_2) < (L_1+1+i)$. In this context, the following theorem is presented.

Theorem 6.1

Given that $T_{\max}(L_1-1, L_2) < T \leq T_{\max}(L_1, L_2)$, $2 \leq L_1+L_2 \leq N_1+N_2$, the optimal control of a GE(n)/GE/1/N queue with two classes of users and

a maximum of $N=N_1+N_2$ controlled packets in the system are given by

$$\lambda(N-n_1-n_2) = \left\{ \begin{array}{l} \lambda_{\max}, \quad \left. \begin{array}{l} n_1 = 0, 1, \dots, L_1-2, \\ n_2 = 0, 1, \dots, L_2 \\ n_1+n_2 = 0, 1, \dots, L_1+L_2-2, \end{array} \right\} \\ \sum_{\substack{m_1=0 \\ (m_1+m_2>0)}}^{L_1-1} \sum_{m_2=0}^{L_2} \left[\frac{[\mu T - m_1 - m_2] \binom{m_1+m_2}{m_1} \hat{x}_1(m_1, m_2) \hat{x}_2(0, m_2)}{[L_1+L_2 - \mu T] \binom{L_1+L_2}{L_1} \hat{x}_1(L_1-1, m_2) \hat{x}_2(0, m_2)} \right] \\ \left[\frac{\mu \tau_2 - \pi \mu \tau_2 [1 - \tau_1]}{\tau_1 \pi} \right] - \frac{\mu \tau_2 [1 - \tau_1]}{\tau_1}, \quad \begin{array}{l} n_1 = L_1-1, \\ n_2 = L_2, \end{array} \\ 0, \quad \left. \begin{array}{l} n_1 = L_1, L_1+1, \dots, N_1, \\ n_2 = L_2, L_2+1, \dots, N_2, \\ n_1+n_2 = L, L+1, \dots, N, \end{array} \right\} \end{array} \right. \quad [6.42]$$

and the maximum throughput is given by

$$F(T) = \mu \left[\frac{\sum_{\substack{n_1=0 \\ (n_1+n_2>0)}}^{L_1} \sum_{n_2=0}^{L_2} \binom{n_1+n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)}{1 + \sum_{\substack{n_1=0 \\ (n_1+n_2>0)}}^{L_1} \sum_{n_2=0}^{L_2} \binom{n_1+n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)} \right] \quad [6.43]$$

Proof

Let us assume that

$$T_{\max}(L_1-1, L_2) < T \leq T_{\max}(L_1, L_2) \quad [6.44]$$

which is true as shown in Lemma 6.3 above, for all L_1+L_2 , $L_1+L_2 = 2, 3, \dots, N$. The condition for the maximum mean time delay in Lemma 6.1 can be written as

$$\begin{aligned} x_1(L_1, L_2)x_2(0, L_2) &\leq -\frac{[N_1+N_2-\mu T]}{[L_1+L_2-\mu T] \binom{L_1+L_2}{L_1}} x_1(N_1, N_2)x_2(0, N_2) - \dots \\ &- \frac{[L_1+L_2+1-\mu T]}{[L_1+L_2-\mu T] \binom{L_1+L_2}{L_1}} \left[x_1(L_1+1, L_2)x_2(0, L_2) + x_1(L_1, L_2+1)x_2(0, L_2+1) \right] \\ &- \frac{[L_1+L_2-1-\mu T]}{[L_1+L_2-\mu T] \binom{L_1+L_2}{L_1}} \left[x_1(L_1-1, L_2)x_2(0, L_2) + x_1(L_1, L_2-1)x_2(0, L_2-1) \right] \\ &- \dots - \frac{[2-\mu T]}{[L_1+L_2-\mu T] \binom{L_1+L_2}{L_1}} \left[x_1(1, 1)x_2(0, 1) + x_1(2, 0)x_2(0, 0) \right] \\ &- \frac{[1-\mu T]}{[L_1+L_2-\mu T] \binom{L_1+L_2}{L_1}} \left[x_1(1, 0)x_2(0, 0) + x_1(0, 1)x_2(0, 1) \right] \end{aligned} \quad [6.45]$$

and can be written as

Equality of the relation [6.48] is satisfied if and only if

$$x_1(n_1, n_2) x_2(0, n_2) =$$

$$\left\{ \begin{array}{ll} 0, & \begin{array}{l} n_1 = L_1, \dots, N_1, \\ n_2 = L_2, \dots, N_2, \\ n_1 + n_2 > L_1 + L_2, \end{array} \\ \sum_{\substack{m_1=0 \\ (m_1+m_2>0)}}^{L_1-1} \sum_{m_2=0}^{L_2} \frac{[\mu T - m_1 - m_2] \binom{m_1+m_2}{m_1} \hat{x}_1(m_1, m_2) \hat{x}_2(0, m_2)}{[L_1 + L_2 - \mu T] \binom{L_1+L_2}{L_1}}, & \begin{array}{l} n_1 = L_1, \\ n_2 = L_2, \end{array} \\ \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2), & \begin{array}{l} n_1 = 0, 1, \dots, L_1, \\ n_2 = 0, 1, \dots, L_2, \\ 0 < n_1 + n_2 < L_1 + L_2, \end{array} \end{array} \right. \quad [6.49]$$

It can be shown that

$$0 < x_1(L_1, L_2) x_2(0, L_2) \leq \hat{x}_1(L_1, L_2) \hat{x}_2(0, L_2) \quad [6.50]$$

can be achieved by substituting [6.49] into [6.50]. It follows that

$$0 < \sum_{\substack{n_1=0 \\ (n_1+n_2>0)}}^{L_1-1} \sum_{n_2=0}^{L_2} \frac{[\mu T - (n_1 + n_2)] \binom{n_1+n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)}{[L_1 + L_2 - \mu T] \binom{L_1+L_2}{L_1}} \leq \hat{x}_1(L_1, L_2) \hat{x}_2(0, L_2) \quad [6.51]$$

which is equivalent to

$$\begin{aligned}
& \frac{\sum_{n_1=0}^{L_1-1} \sum_{n_2=0}^{L_2} \binom{n_1+n_2}{n_1} [n_1+n_2] \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)}{(n_1+n_2 > 0)} < T \leq \\
& \mu \sum_{n_1=0}^{L_1-1} \sum_{n_2=0}^{L_2} \binom{n_1+n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2) \\
& (n_1+n_2 > 0) \\
& \frac{\sum_{n_1=0}^{L_1-1} \sum_{n_2=0}^{L_2} \binom{n_1+n_2}{n_1} [n_1+n_2] \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)}{(n_1+n_2 > 0)} + \\
& \mu \left[\sum_{n_1=0}^{L_1-1} \sum_{n_2=0}^{L_2} \binom{n_1+n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2) + \binom{L_1+L_2}{L_1} \hat{x}_1(L_1, L_2) \hat{x}_2(0, L_2) \right] \\
& (n_1+n_2 > 0) \\
& \frac{[L_1+L_2] \binom{L_1+L_2}{L_1} \hat{x}_1(L_1, L_2) \hat{x}_2(0, L_2)}{\mu \left[\sum_{n_1=0}^{L_1-1} \sum_{n_2=0}^{L_2} \binom{n_1+n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2) + \binom{L_1+L_2}{L_1} \hat{x}_1(L_1, L_2) \hat{x}_2(0, L_2) \right]} \\
& (n_1+n_2 > 0)
\end{aligned} \tag{6.52}$$

or from Lemma 6.3, [6.52] can be written as in [6.44]. Then it can be shown from [6.49] that Theorem 6.1 is true as follows. Since from [6.49]

$$x_1(1, 0) x_2(0, 0) = \hat{x}_1(1, 0) \hat{x}_2(0, 0) \tag{6.53}$$

and from [6.19], [6.20], [6.29] and [6.30]

$$\frac{\pi\lambda(N)\tau_1\tau_2}{\mu\tau_2+\lambda(N-1)\tau_1(1-\tau_2)} = \frac{\pi\lambda_{\max}\tau_1\tau_2}{\mu\tau_2+\lambda_{\max}\tau_1(1-\tau_2)} \quad [6.54]$$

It can be shown that

$$\lambda(N) = \lambda_{\max} \quad [6.55]$$

It follows from [6.49] that

$$x_1(n_1, n_2)x_2(0, n_2) = \hat{x}_1(n_1, n_2)\hat{x}_2(0, n_2), \quad 1 \leq n_1 + n_2 \leq L_1 + L_2 - 1 \quad [6.56]$$

and from the above relationship, it can be similarly shown that

$$\lambda(N - n_1 - n_2) = \lambda_{\max}, \quad n_1 + n_2 = 1, 2, \dots, L_1 + L_2 - 2. \quad [6.57]$$

When the system has achieved maximum capacity, it is implied that

$$x_1(L_1, L_2)x_2(0, L_2) = \sum_{\substack{n_1=0 \\ (n_1+n_2>0)}}^{L_1-1} \sum_{n_2=0}^{L_2} \frac{[\mu T - (n_1 + n_2)] \binom{n_1 + n_2}{n_1} \hat{x}_1(n_1, n_2)\hat{x}_2(0, n_2)}{[L_1 + L_2 - \mu T] \binom{L_1 + L_2}{L_1}}, \quad [6.58]$$

which is equivalent to

$$x_1(L_1, L_2)x_2(0, L_2) = \hat{x}_1(L_1 - 1, L_2)\hat{x}_2(0, L_2) \pi \left[\frac{\mu\tau_2[1-\tau_1] + \lambda(N - L_1 + 1 - n_2)\tau_1}{\mu\tau_2 - \pi\mu\tau_2[1-\tau_1]} \right] \quad [6.59]$$

and thus it follows that [6.42] holds for $n_1 = L_1 - 1$ and $n_2 = L_2$. Note that, since the queueing centre 0 will contain at all times at least $N - L_1 - L_2$ packets, the optimum control can also be achieved by a control scheme using a total of $N = L_1 + L_2$ packets (i.e. fixed window scheme). It is required that

$$\lambda(L - n_1 - n_2) = \begin{cases} \lambda_{\max}, & \begin{aligned} & n_1 = 0, 1, \dots, L_1 - 2, \\ & n_2 = 0, 1, \dots, L_2, \\ & n_1 + n_2 < L_1 + L_2 - 1, \end{aligned} \\ \sum_{m_1=0}^{L_1-1} \sum_{\substack{m_2=0 \\ (m_1+m_2 > 0)}}^{L_2} \left[\frac{[\mu T - m_1 + m_2] \binom{m_1 + m_2}{m_1} \hat{x}_1(m_1, m_2) \hat{x}_2(0, m_2)}{[L_1 + L_2 - \mu T] \binom{L_1 + L_2}{L_1} \hat{x}_1(L_1 - 1, m_2) \hat{x}_2(0, m_2)} \right] \\ \left[\frac{\mu \tau_2 - \pi [\mu \tau_2 (1 - \tau_1)]}{\tau_1 \pi} \right] - \frac{\mu \tau_2 (1 - \tau_1)}{\tau_1}, & \begin{aligned} & n_1 = L_1 - 1, \\ & n_2 = L_2, \end{aligned} \end{cases} \quad [6.60]$$

Therefore from [6.42] it can be seen that the optimal flow control of a GE/GE/1 queue is a sliding (variable) window type mechanism which is known in control theory as a bang-bang control, that is the optimum offered load is as follows:

- i. equal to the maximum input rate $\lambda_{\max}$ when the number of packets is less than $L_1 + L_2 - 1$,
- ii. less than the maximum input rate when the number of packets is equal to $L_1 + L_2 - 1$, and
- ii. equal to zero when the system is full.

The window size can be derived from the maximum time delay of the system via the following corollary.

Corollary 6.1

If $T = T_{\max}(L)$, then the throughput time delay function and the maximum time delay are given by

$$F[T_{\max}(L_1, L_2)] = \frac{\sum_{n_1=0}^{L_1} \sum_{n_2=0}^{L_2} \mu \binom{n_1+n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)}{1 + \sum_{n_1=0}^{L_1} \sum_{n_2=0}^{L_2} \binom{n_1+n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)}, \quad [6.61]$$

and

$$T_{\max}(L_1, L_2) = \frac{\sum_{n_1=0}^{L_1} \sum_{n_2=0}^{L_2} \binom{n_1+n_2}{n_1} [n_1+n_2] \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)}{\sum_{n_1=0}^{L_1} \sum_{n_2=0}^{L_2} \mu \binom{n_1+n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)}, \quad [6.62]$$

respectively.

Proof

Since $T = T_{\max}(L_1, L_2)$, it follows that

$$\begin{aligned}
 x_1(L_1, L_2) x_2(0, L_2) &= \sum_{\substack{n_1=0 \\ (n_1+n_2>0)}}^{L_1-1} \sum_{n_2=0}^{L_2} \frac{[\mu T - (n_1 + n_2)] \binom{n_1 + n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)}{[L_1 + L_2 - \mu T] \binom{L_1 + L_2}{L_1}} \\
 &= \hat{x}_1(L_1, L_2) \hat{x}_2(0, L_2) \quad [6.63]
 \end{aligned}$$

and thus substituting [6.49] into [6.1] and [6.2] gives [6.61] and [6.62].

Numerical Examples

In this section numerical examples are presented to illustrate how critically the throughput time delay function, $F[T_{\max}(L_1, L_2)]$, is affected by the distributional form of the interarrival time and service time distributions. The GE distributional model is chosen to approximate general distributions with known first two moments.

The maximum throughput as a function of the control are depicted in Figures 64-67 for C_μ^2 , $C_{\lambda_1}^2$ and $C_{\lambda_2}^2$ varying from 0.5 to 5.0, $c_1 = 0.3$, $c_2 = 0.6$ and $\mu = 1$. It is noted that the GE-type solutions for the maximum throughput as a function of the control provides lower bounds (pessimistic) when C_μ^2 , $C_{\lambda_1}^2$ and $C_{\lambda_2}^2 = 5$ and upper bounds (optimistic) when C_μ^2 , $C_{\lambda_1}^2$ and $C_{\lambda_2}^2 = 0.5$. The maximum throughput as a function of the control for values of $c_1 + c_2 = 0.5, 0.7$ and 0.9 with C_μ^2 , $C_{\lambda_1}^2$ and

$C_{\lambda_2}^2 = 5.0$, $\mu = 1$ are depicted in Figures 67–69. It can be observed that the GE(n)/GE/1/N solution predicts the increasing optimistic values as expected for throughput, as c_1+c_2 moves between 0.5 and 0.9. Also as expected, the maximum time delay as a function of the control provides lower bounds (pessimistic) when C_{μ}^2 , $C_{\lambda_1}^2$ and $C_{\lambda_2}^2 = 0.5$ and upper bounds (optimistic) when C_{μ}^2 , $C_{\lambda_1}^2$ and $C_{\lambda_2}^2 = 5$ (see Figures 70–73). The maximum time delay as a function of the control for values of $c_1+c_2 = 0.5$, 0.7 and 0.9 with C_{μ}^2 , $C_{\lambda_1}^2$ and $C_{\lambda_2}^2 = 5.0$, $\mu = 1$ are depicted in Figures 73–75. It can be observed that the GE(n)/GE/1/N solution predicts the increasing optimistic values as expected, for throughput as c_1+c_2 moves between 0.5 and 0.9.

In order to display the relationship between the maximum time delay constraint and the maximum throughput, it is necessary to show their dependence on the control. To do so, first, the control of user class 1 is parameterised and the control of user class 2 is increased. Secondly, the control of user class 2 is parameterised and that of user class 1 is increased. The resulting throughput–time delay trade-off are shown in Figures 76 and 77 respectively for $C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 1$. The optimal control is obtained by comparing these curves and choosing the control which corresponds to a maximum throughput for a given time delay constraint T . By superimposing the curves in Figures 76 and 77, it can be seen that the optimal control allows near window sizes for the two users. If $c_1 < c_2$, one user class 2 packet should enter the system first. If the time delay constraint has not yet been achieved, one user class 1 packet has to enter the system. This process is repeated alternating between user class 1 and 2 until the

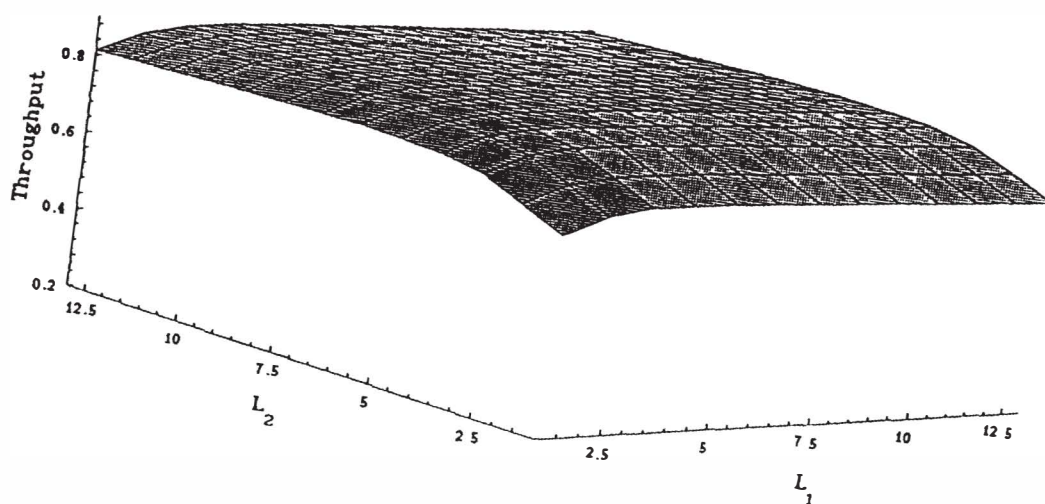


Figure 64 : Throughput versus Control
 $(C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 0.5, c_1 = 0.3, c_2 = 0.6, \mu = 1)$

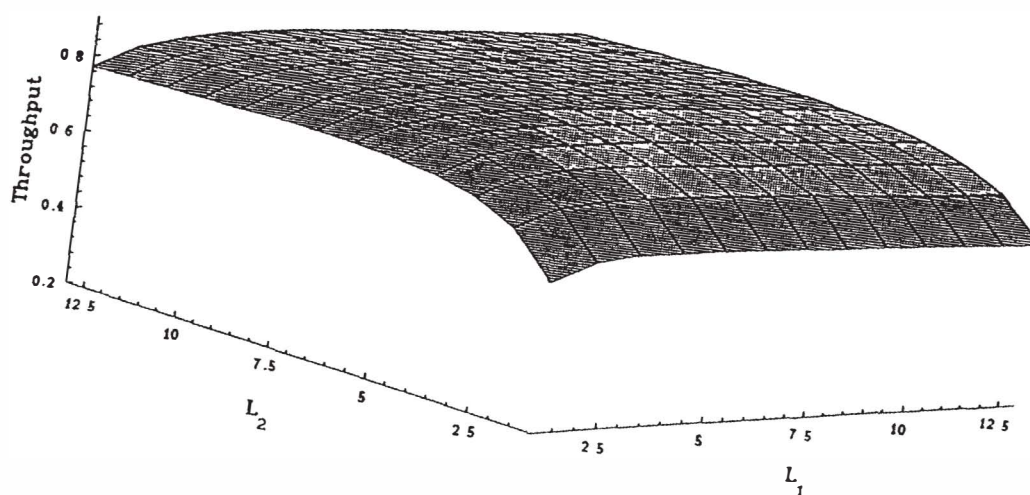


Figure 65 : Throughput versus Control
 $(C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 1.0, c_1 = 0.3, c_2 = 0.6, \mu = 1)$

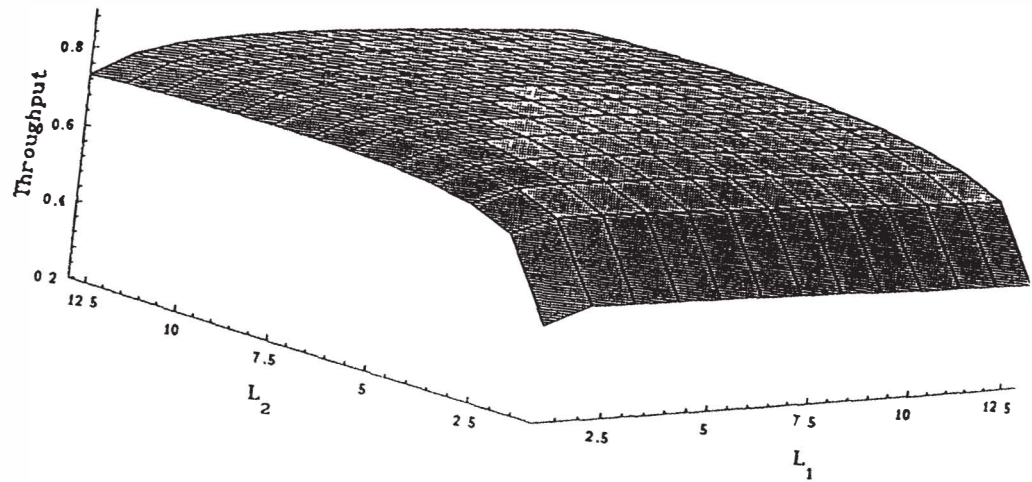


Figure 66 : Throughput versus Control
 $(C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 2.0, c_1 = 0.3, c_2 = 0.6, \mu = 1)$

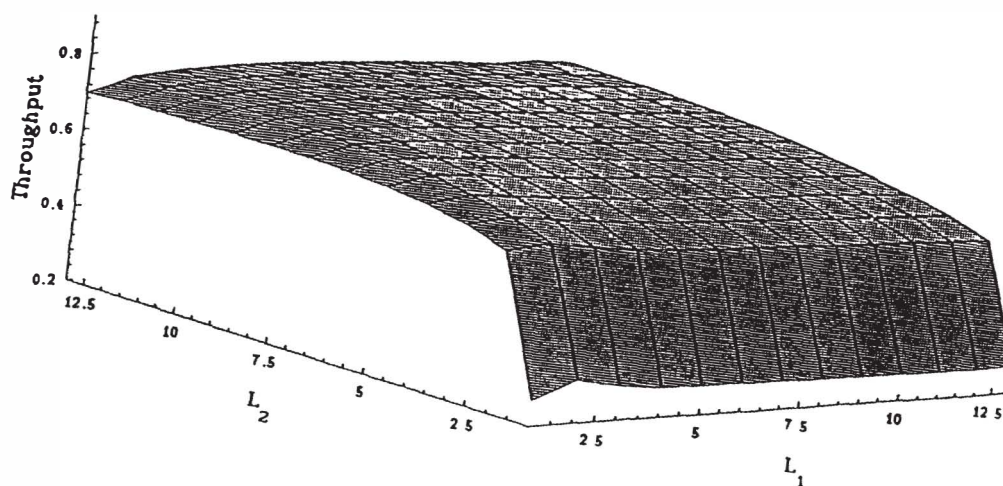


Figure 67 : Throughput versus Control
 $(C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 5.0, c_1 = 0.3, c_2 = 0.6, \mu = 1)$

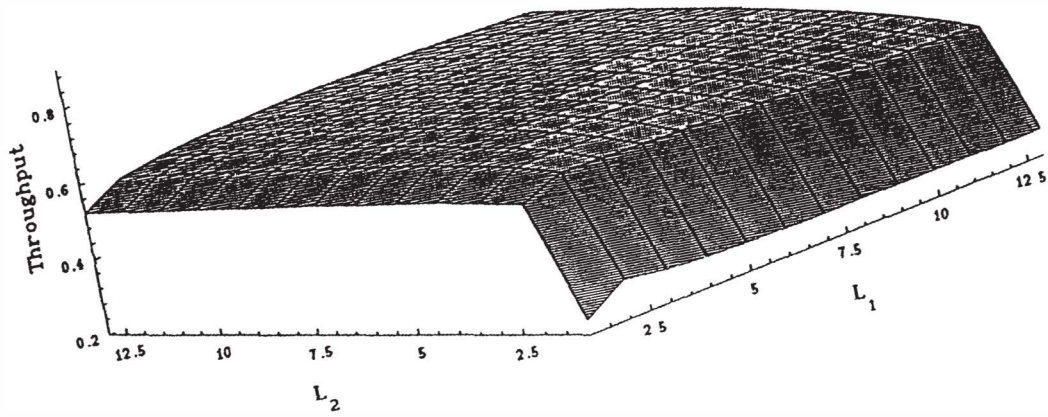


Figure 68 : Throughput versus Control
 $(C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 5.0, c_1 = 0.3, c_2 = 0.4, \mu = 1)$

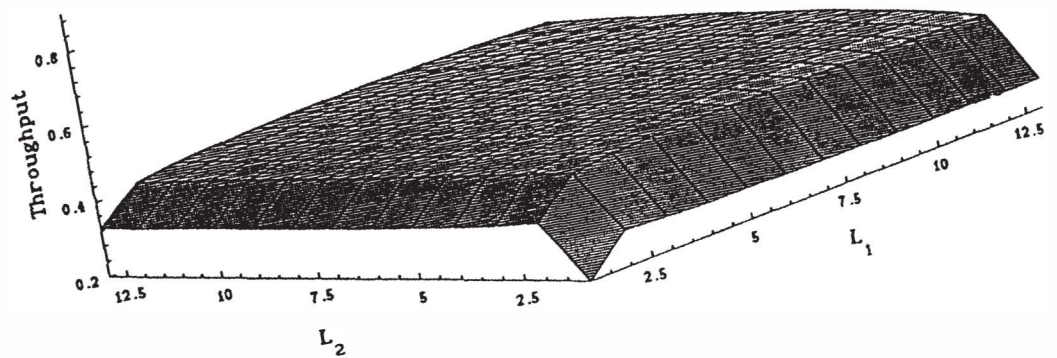


Figure 69 : Throughput versus Control
 $(C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 5.0, c_1 = 0.3, c_2 = 0.2, \mu = 1)$

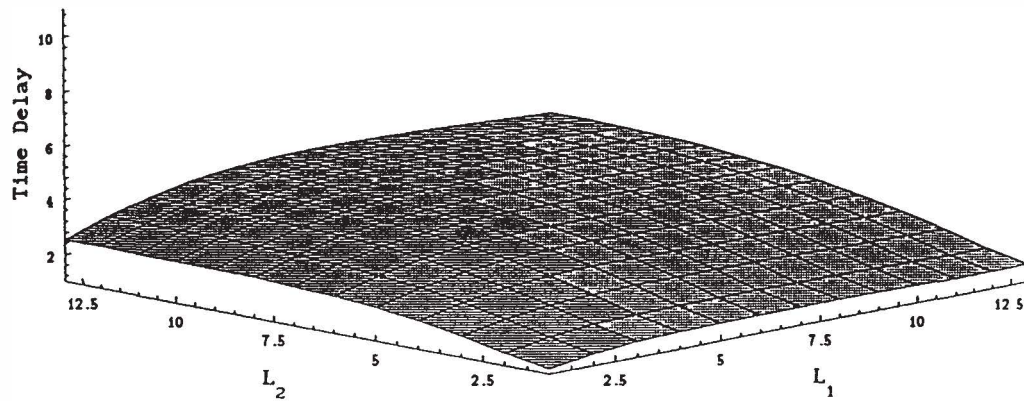


Figure 70 : Time Delay versus Control

$$(C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 0.5, c_1 = 0.3, c_2 = 0.6, \mu = 1)$$

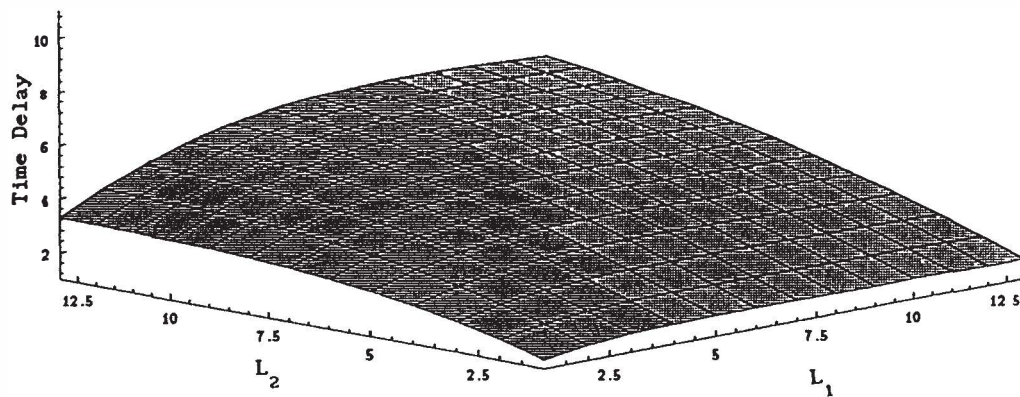


Figure 71 : Time Delay versus Control

$$(C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 1.0, c_1 = 0.3, c_2 = 0.6, \mu = 1)$$

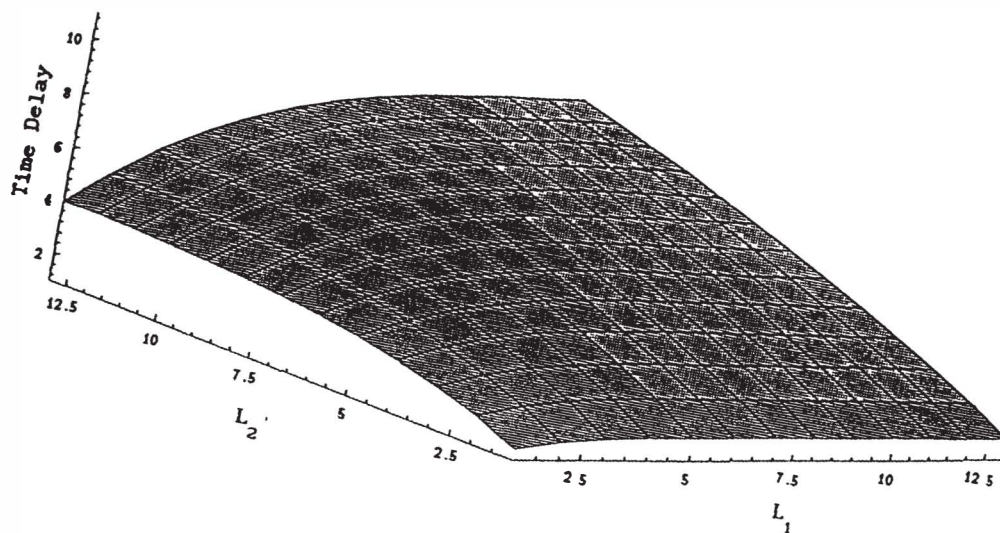


Figure 72 : Time Delay versus Control
 $(C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 2.0, c_1 = 0.3, c_2 = 0.6, \mu = 1)$

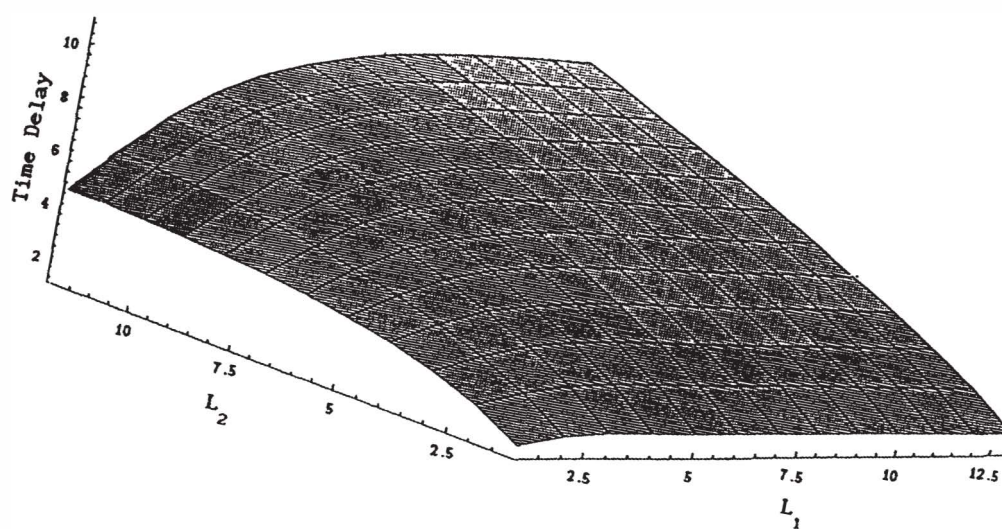


Figure 73 : Time Delay versus Control
 $(C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 5.0, c_1 = 0.3, c_2 = 0.6, \mu = 1)$.

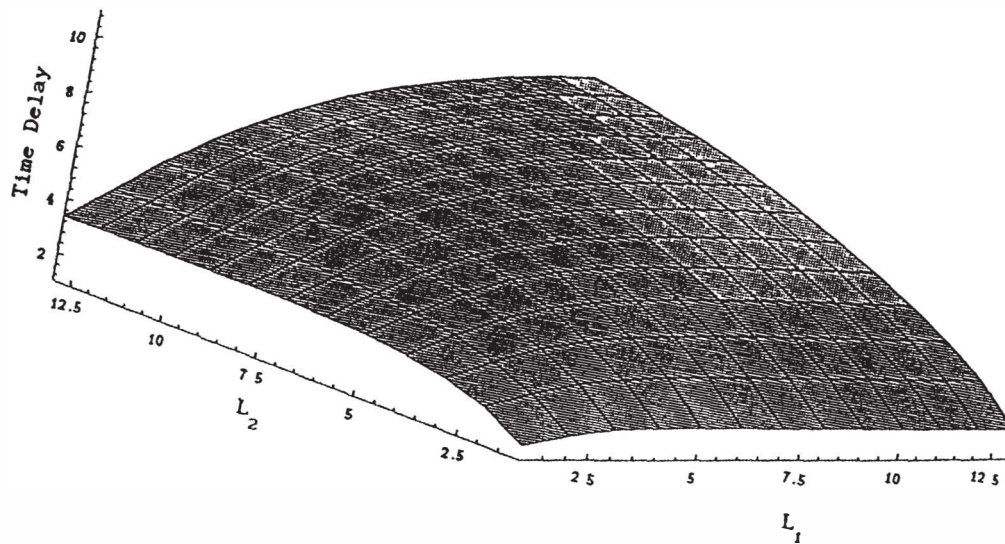


Figure 74 : Time Delay versus Control
 $(C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 5.0, c_1 = 0.3, c_2 = 0.4, \mu = 1)$

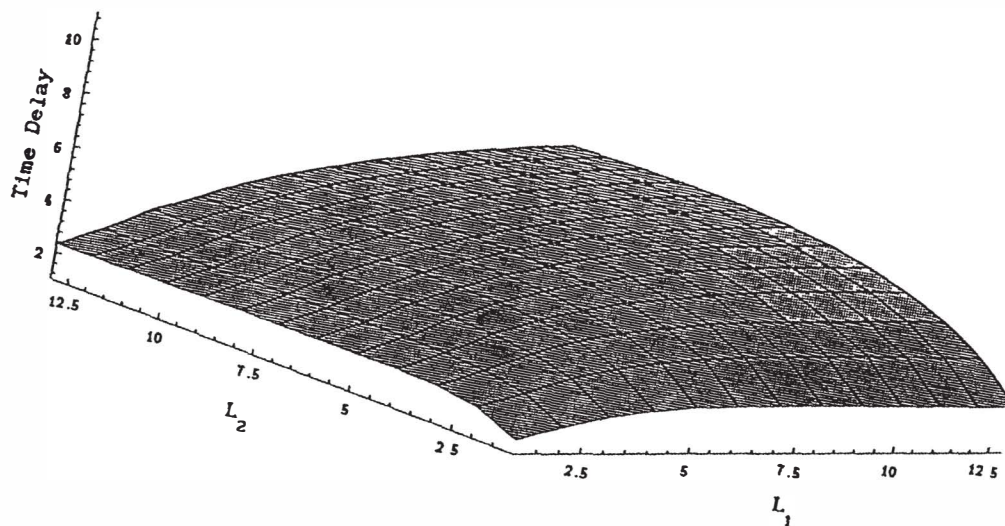


Figure 75 : Time Delay versus Control
 $(C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 5.0, c_1 = 0.3, c_2 = 0.2, \mu = 1)$

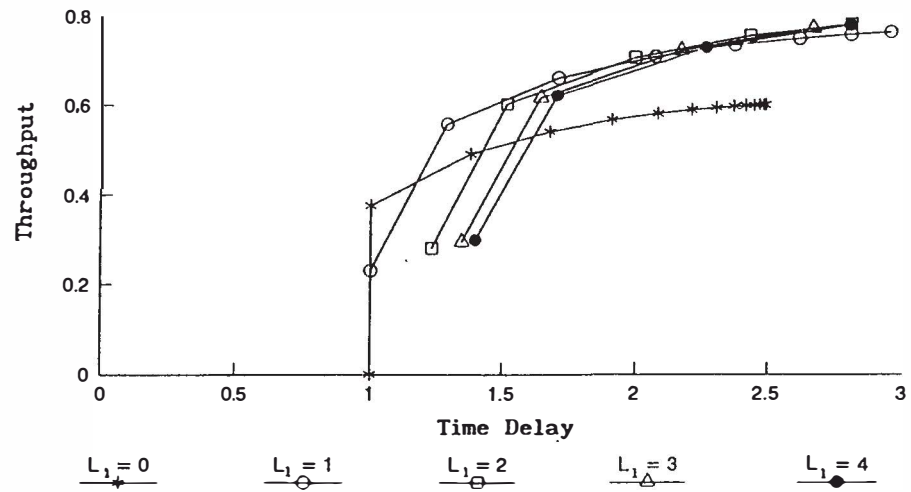


Figure 76 : Throughput versus Time Delay

$$(L_1 \text{ is fixed, } C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 1.0)$$

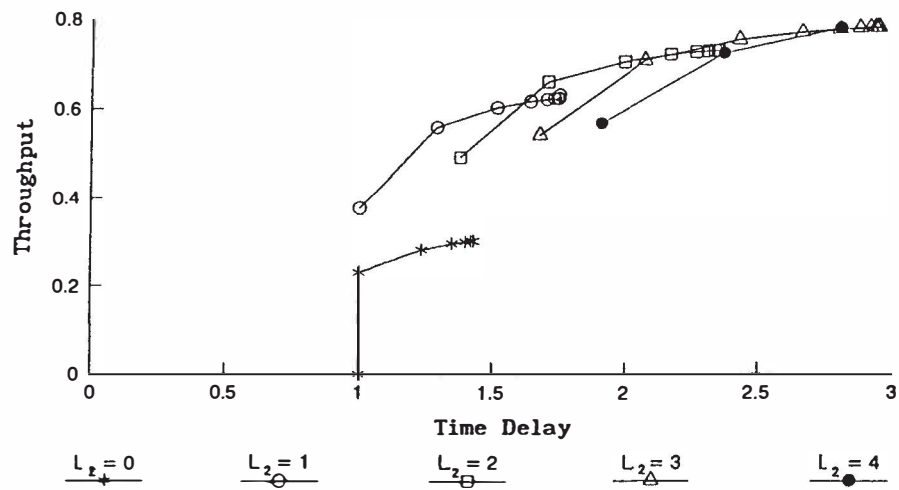


Figure 77 : Throughput versus Time Delay

$$(L_2 \text{ is fixed, } C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 1.0)$$

time delay constraint is achieved. It is important to note that only the order relation between c_1 and c_2 is needed in the determination of the optimal window sizes for the two users. This does not appear to be intuitive from a system point of view, since the resulting optimal strategy gives equal access to the system resources even though the users have different rates. From a fairness point of view, however, the result is intuitively pleasing since a global objective should not favour any particular type of users.

Summary

In this chapter, the decentralised optimal flow control of a G/G/1 queue with K classes of users approximated by a GE/GE/1 queue, when the first two moments are known, has been studied. The optimal control that maximises the average throughput subject to a bounded time delay constraint has been shown to be a state dependent window type mechanism. The maximum number of packets in transit within the system (i.e. sliding window size) corresponding to a maximum throughput (when optimally controlled), can be determined from a pre-assigned upper bound on the average time delay. The direct dependence of the maximum throughput on the average time delay is also examined. Numerical examples provide useful information on how critically the optimal throughput is affected by the distributional form of the interarrival and service times.

The results described in this work provide a flexible tool in designing or upgrading computer communication networks with non exponential data flows. The analysis of the optimal flow control of a GE/GE/1 queue reconfirms (Hsiao and Lazar, 1984) that 'the optimal

control under the global objectives does not favour any individual user, and depends only on the relative order of their packet generation rates'.

CHAPTER VII

DECENTRALISED OPTIMAL FLOW CONTROL OF MULTISERVER GENERAL QUEUEING NETWORKS WITH MULTICLASS USERS

Introduction

This chapter is concerned with the decentralised optimal flow control of computer communication networks with general packet interarrival and transmission time distributions. As mentioned earlier, flow control is intended to provide fast and reliable communication with efficient allocation of network resources. The performance of a network can be greatly degraded by high utilisation links (bottlenecks). One purpose of flow control is to make efficient use of network resources by limiting or rerouting traffic, thus avoiding bottlenecks.

Recent work on optimal flow control in computer communication network with multiclass users has been focused on single transmission link (Hsiao and Lazar, 1984; 1985). The authors studied the decentralised local flow control in computer communication networks with exponential packet interarrival and transmission patterns, and two classes of users sharing a high utilisation link. The decentralised flow control of the link is shown to be a pair of window type mechanisms where the optimal window sizes depend on the maximum packet generation rate of the two users, the average service rate and the maximum tolerable average time delay. However, it is also essential to extend these results on multiple transmission links in order to investigate how critically the flow control strategy and the optimal throughput are affected by the distributional form of packet

interarrival and transmission patterns in computer communication networks.

Using Hsiao and Lazar (1984) optimality criterion, this chapter examines the decentralised optimal flow of high utilisation links in computer communication networks. Since one of the main design requirements for communication protocols is the acknowledgment of packets reaching their destination, a general closed queueing network is used as a model for studying the decentralised optimal flow control problem (Lam and Lien, 1981).

The decentralised flow control problem considered here assumes that resources are completely shared. For simplicity, it is assumed that only two classes of users share the service of the links on a FCFS discipline. User packets queue up in a common buffer to wait for transmission. The optimal decentralised flow control is obtained for a larger class of control mechanisms, which incorporates local information on the number of outstanding packets. The optimality criterion also allows explicit comparison of the throughput time-delay trade-off.

In view of the multiplicity of user classes, there exist naturally more than one objective to be considered (Lazar, 1984). Again as in Chapter VI, this chapter considers the global objective of maximizing the average throughput subject to a bound on the average time delay over all the users sharing the links.

This chapter is divided into the following sections. The model for the bottleneck decentralised flow control is presented. Relevant optimisation problem is then defined using the global objective function. The optimal control of a $G/G/\infty$ finite capacity queue at

equilibrium which is a special case of the G/G/c finite capacity queue subject to a bounded time delay is then derived followed by the derivation of the optimal flow control for the G/G/c finite capacity queue. Numerical examples will then be provided followed by a summary of this chapter.

The Mathematical Model

The model for the decentralised control problem is shown in Figure 78. In the following, the bottleneck which consists of c (≥ 2) parallel homogeneous servers is modeled as a general server with load dependent rates $\mu(\underline{n})$, $\underline{n} = (n_1, n_2, \dots, n_K)$, constant coefficient of variation C_μ and a FCFS serving discipline. K classes of users are considered. Packets belonging to either class queue up in a common buffer to wait for transmission via the bottleneck links. User class i , $i = 1, 2, \dots, K$, packets are generated at a maximum rate c_i packets per second. The decentralised optimal flow control problem is formulated as followed. Each user controls his rate of packet generation using information of the number of outstanding packets (i.e via acknowledgment or otherwise). Information on the state of the other user class is not available. It is assumed that the number of class i packets is N_i where N_i are arbitrarily large integers.

The model in Figure 78 is a general closed queueing network which has been studied in Chapter V. Here, $\underline{n} = (n_1, n_2, \dots, n_K)$ is the state of the server (i.e link), n_i being the number class i packets in the queue, respectively. $\lambda_i(N_i - n_i)$ is the corresponding controlled packet generation rates with n_i outstanding packets. Centre 0 can be viewed as a G/G/c queue with finite buffer size N denoted by $G(n)/G(n)/c/N$ where $n = n_1 + n_2$ and $N = N_1 + N_2 + \dots + N_K$.

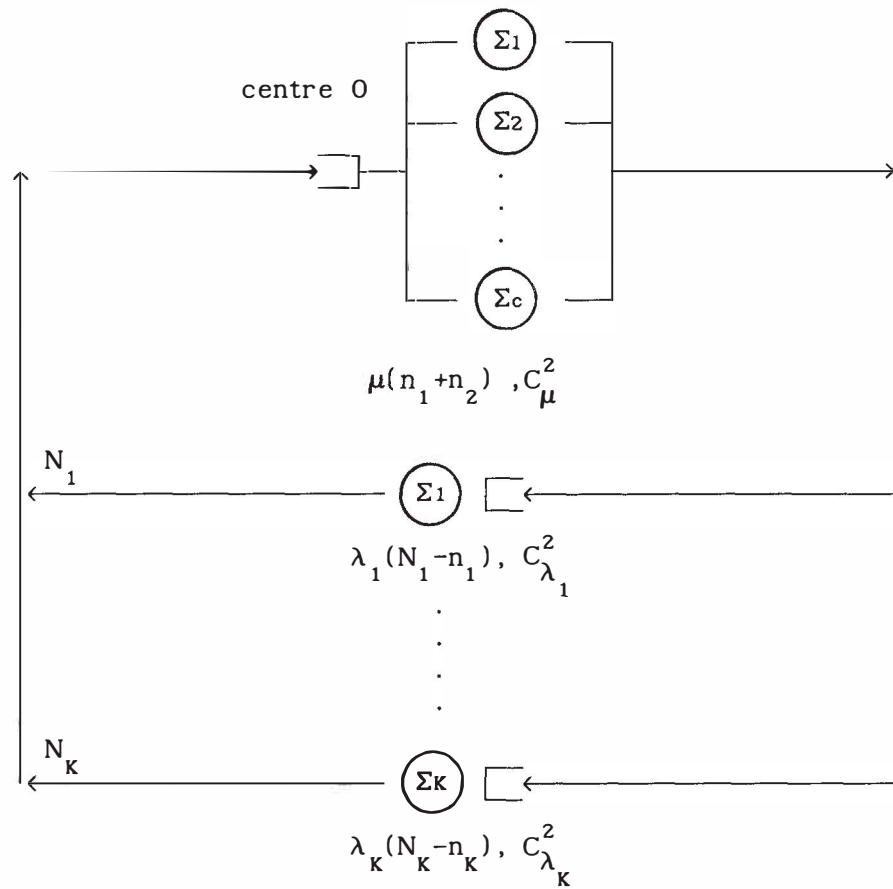


Figure 78 : Queueing Model for Bottleneck Decentralised Flow Control

The mean throughput and the mean time delay of the bottleneck link of Figure 78 are given by

$$\langle \chi \rangle N = \sum_{n_1=0}^{N_1} \dots \sum_{n_K=0}^{N_K} \mu(n_1+n_2+\dots+n_K) p_N(\underline{n}) \quad [7.1]$$

($n_1+\dots+n_K > 0$)

and by Little's law

$$\langle \tau \rangle_N = \frac{\sum_{n_1=0}^{N_1} \dots \sum_{n_K=0}^{N_K} (n_1 + n_2 + \dots + n_K) p_N(\underline{n})}{\sum_{n_1=0}^{N_1} \dots \sum_{n_K=0}^{N_K} \mu(n_1 + n_2 + \dots + n_K) p_N(\underline{n})}, \quad [7.2]$$

(n₁ + ... + n_K > 0)

respectively, where $p_N(\underline{n})$ is the corresponding joint queue length distribution of the bottleneck link of Figure 78 and can be analogously expressed by

$$p_N(\underline{n}) = p_N(0) \binom{n_1 + n_2 + \dots + n_K}{n_1} \left[\prod_{j=0}^{n_1-1} \frac{a_1(j, n_2, n_3, \dots, n_K)}{b_1(j+1, n_2, n_3, \dots, n_K)} \right] \binom{n_2 + n_3 + \dots + n_K}{n_2} \left[\prod_{j=0}^{n_2-1} \frac{a_2(0, j, n_3, \dots, n_K)}{b_2(0, j+1, n_3, \dots, n_K)} \right] \dots \left[\prod_{j=0}^{n_K-1} \frac{a_K(0, 0, \dots, 0, j)}{b_K(0, 0, \dots, 0, j+1)} \right], \quad [7.3]$$

$n_i = 1, 2, \dots, N_i,$
 $i = 1, 2, \dots, K,$
 $n_1 + n_2 + \dots + n_K \neq 0,$

and

$$p_N(0) = \left[1 + \sum_{n_1=0}^{N_1} \dots \sum_{n_K=0}^{N_K} p_N(\underline{n}) \right]^{-1}, \quad [7.4]$$

(n₁ + ... + n_K > 0)

where $a_i(\underline{n})$ and $b_i(\underline{n})$ ($i = 1, 2, \dots, K$) are the 'up' and 'down' rates. Due to the complexity involved in writing down $a_i(\underline{n})$ and $b_i(\underline{n})$ for K classes of users, we will assume that only two classes (i.e. $K = 2$) of users share the service of the link on a FCFS discipline. The

solutions can easily be extended to more than two classes of users. The joint queue length distribution of the bottleneck link of Figure 78 for two classes of users can be expressed by

$$p_N(n_1, n_2) = \binom{n_1+n_2}{n_1} \left[\prod_{j=0}^{n_1-1} \frac{a_1(j, n_2)}{b_1(j+1, n_2)} \right] \left[\prod_{j=0}^{n_2-1} \frac{a_2(0, j)}{b_2(0, j+1)} \right] p_N(0, 0), \quad [7.5]$$

and

$$p_N(0, 0) = \left[1 + \sum_{\substack{n_1=0 \\ (n_1+n_2>0)}}^{N_1} \sum_{n_2=0}^{N_2} \binom{n_1+n_2}{n_1} \left[\prod_{j=0}^{n_1-1} \frac{a_1(j, n_2)}{b_1(j+1, n_2)} \right] \left[\prod_{j=0}^{n_2-1} \frac{a_2(0, j)}{b_2(0, j+1)} \right] \right]^{-1}, \quad [7.6]$$

where $\prod_{j=0}^{-1} = 1$ and the parameters $a_1(n_1, n_2)$, $b_1(n_1, n_2)$, $a_2(0, n_2)$ and $b_2(0, n_2)$ represent 'up' and 'down' flow coefficients (see Figure 79) given by

$$a_1(n_1, n_2) = \begin{cases} \pi\lambda(N)\tau_1, & n_1 + n_2 = 0, \\ \pi t, & \left. \begin{array}{l} n_1 = 0, 1, \dots, c-2, \\ n_2 = 0, 1, \dots, c-2, \end{array} \right\} 0 < n_1 + n_2 \leq c-2, \\ \pi\tau_2 e, & \left. \begin{array}{l} n_1 = 0, 1, \dots, c-1, \\ n_2 = 0, 1, \dots, c-1, \end{array} \right\} n_1 + n_2 = c-1, \\ \pi q, & \left. \begin{array}{l} n_1 = 0, 1, \dots, N_1-1, \\ n_2 = 0, 1, \dots, N_2, \end{array} \right\} n_1 + n_2 \geq c, \end{cases} \quad [7.7]$$

$$a_2(0, n_2) = \begin{cases} [1-\pi]\lambda(N)\tau_1, & n_2 = 0, \\ [1-\pi]t, & n_2 = 1, 2, \dots, c-2, \\ [1-\pi]\tau_2 e, & n_2 = c-1, \\ [1-\pi]q, & n_2 = c, c+1, \dots, N_2-1, \end{cases} \quad [7.9]$$

$$b_2(0, n_2) = \begin{cases} n_2 \mu(n_1 + n_2) [\tau_1 + \tau_2 - \tau_1 \tau_2] & n_2 = 1, 2, \dots, c-1, \\ f, & n_2 = c, c+1, \dots, N_2-1 \\ \frac{f \tau_1 \tau_2 [c \mu(n_1 + n_2) - \pi \lambda(N - n_2)]}{f - \pi q} & n_2 = N_2, \end{cases} \quad [7.10]$$

$$t = [n_1 + n_2] \mu(n_1 + n_2) \tau_2 [1 - \tau_1] + \lambda(N - n_1 - n_2) \tau_1,$$

$$e = [c-1] \mu(n_1 + n_2) \tau_2 [1 - \tau_1] + \lambda(N - n_1 - n_2) \tau_1,$$

$$q = c \mu(n_1 + n_2) \tau_2 [1 - \tau_1] + \lambda(N - n_1 - n_2) \tau_1,$$

$$f = c \mu(n_1 + n_2) \tau_2 + \tau_1 [1 - \tau_2] \lambda(N - n_1 - n_2),$$

$$\pi = \frac{\lambda_1 (N_1 - n_1)}{\lambda(N - n_1 - n_2)}, \quad \tau_1 = \frac{2}{(C_a^2 + 1)}, \quad \tau_2 = \frac{2}{(C_\mu^2 + 1)}, \text{ and}$$

$$C_a^2 = \frac{\lambda(N - n_1 - n_2)}{\frac{\lambda_1 (N_1 - n_1)}{C_{\lambda_1}^2 + 1} + \frac{\lambda_2 (N_2 - n_2)}{C_{\lambda_2}^2 + 1}} - 1.$$

Note that the joint queue length distribution of the GE(n)/GE(n)/1/N queue can be interpreted as the joint steady state probability of centre 0 containing n packets. The versatile form of the joint queue

length distribution is used in the next section to formulate the optimality criterion. For simplicity we focus on class 1 user packets and the results for class 2 user packets are symmetrical to those obtained for class 1 user packets.

The Optimisation Criterion

Consider the control parameters of the general cyclic queue of Figure 78 for $K = 2$. Let us denote $\lambda_1 = \{\lambda_1(n_1)\}$ for $n_1 = 1, 2, \dots, N_1$, and $\lambda_2 = \{\lambda_2(n_2)\}$ for $n_2 = 1, 2, \dots, N_2$, as a set of control parameters which satisfy the peak constraints

$$0 \leq \lambda_1(n_1) \leq c_1 \quad [7.11]$$

$$0 \leq \lambda_2(n_2) \leq c_2 \quad [7.12]$$

for all n_1 and n_2 , for $n_1 = 1, 2, \dots, N_1$, $n_2 = 1, 2, \dots, N_2$, where c_1 and $c_2 \in \mathbb{R}_+$ are constant called admissible. The maximum input rate rates c_1 and c_2 offered by the source nodes are assumed to be known. The set of control parameters $\lambda_1 = \{\lambda_1(n_1)\}$ for $n_1 = 1, 2, \dots, N_1$, and $\lambda_2 = \{\lambda_2(n_2)\}$ for $n_2 = 1, 2, \dots, N_2$, is said to be optimum over the class of admissible controls for a given time delay $T \in \mathbb{R}_+$, if it achieves the maximum

$$F(T) = \max \langle \chi \rangle N \quad [7.13]$$

subject to

$$\langle \tau \rangle N \leq T \quad [7.14]$$

Here $F(T)$ is called the throughput time delay function. For the sake of simplicity, in what follows only homogeneous service times will be considered (i.e. $\mu(n_1+n_2) = \mu$ for all n_1 and n_2 ; $0 \leq n_1 \leq N_1$, $0 \leq n_2 \leq N_2$). Let us denote the expressions

$$x_1(n_1, n_2) = \prod_{j=0}^{n_1-1} \frac{a_1(j, n_2)}{b_1(j+1, n_2)}, \quad [7.15]$$

and

$$x_2(0, n_2) = \prod_{j=0}^{n_2-1} \frac{a_2(0, j)}{b_2(0, j+1)}, \quad [7.16]$$

or can be written in a step recursion as

$$x_1(n_1, n_2) = \frac{a_1(n_1-1, n_2)}{b_1(n_1, n_2)} x_1(n_1-1, n_2), \quad \begin{array}{l} n_1 = 1, 2, \dots, N_1, \\ n_2 = 0, 1, \dots, N_2, \end{array} \quad [7.17]$$

and

$$x_2(0, n_2) = \frac{a_2(0, n_2-1)}{b_2(0, n_2)} x_2(0, n_2-1), \quad n_2 = 1, 2, \dots, N_2, \quad [7.18]$$

where

$$x_1(0, n_2) = x_2(0, 0) = 1, \quad n_2 = 0, 1, \dots, N_2. \quad [7.19]$$

Hence for a $GE(n)/GE/c/N$ queue we have

$$x_1(n_1, n_2) =$$

$$\left\{ \begin{array}{l} \pi \left[\frac{\lambda(N)\tau_1}{n_1 \mu[\tau_1 + \tau_2 - \tau_1 \tau_2]} \right], \quad n_1 = 1, \quad n_2 = 0, \\ \\ x_1(n_1 - 1, n_2) \pi \left[\frac{[n_1 + n_2 - 1] \mu \tau_2 (1 - \tau_1) + \lambda(N - n_1 - n_2 + 1) \tau_1}{[n_1 + n_2] \mu [\tau_1 + \tau_2 - \tau_1 \tau_2]} \right], \\ \quad \left. \begin{array}{l} n_1 = 1, 2, \dots, c-1, \\ n_2 = 0, 1, \dots, c-2, \end{array} \right\} 1 < n_1 + n_2 < c, \\ \\ x_1(n_1 - 1, n_2) \pi \tau_2 \left[\frac{[c-1] \mu \tau_2 [1 - \tau_1] + \lambda(N - n_1 - n_2 + 1) \tau_1}{c \mu \tau_2 + \lambda(N - n_1 - n_2) \tau_1 [1 - \tau_2]} \right], \\ \quad \left. \begin{array}{l} n_1 = 1, 2, \dots, c, \\ n_2 = 0, 1, \dots, c-1, \end{array} \right\} n_1 + n_2 = c, \\ \\ x_1(n_1 - 1, n_2) \pi \left[\frac{c \mu \tau_2 [1 - \tau_1] + \lambda(N - n_1 - n_2 + 1) \tau_1}{c \mu \tau_2 + \lambda(N - n_1 - n_2) \tau_1 [1 - \tau_2]} \right], \quad [7.20] \\ \quad \left. \begin{array}{l} n_1 = 1, 2, \dots, N_1 - 1, \\ n_2 = 0, 1, \dots, N_2, \end{array} \right\} n_1 + n_2 > c, \\ \\ x_1(n_1 - 1, n_2) \pi \left[\frac{c \mu \tau_2 [1 - \tau_1] + \lambda(N - n_1 - n_2 + 1) \tau_1}{c \mu \tau_2 + \lambda(N - n_1 - n_2) \tau_1 [1 - \tau_2]} \right] \\ \frac{c \mu \tau_2 + \lambda(N - n_1 - n_2) \tau_1 [1 - \tau_2] - [1 - \pi] [c \mu \tau_2 [1 - \tau_1] + \lambda(N - n_1 - n_2) \tau_1]}{\tau_1 \tau_2 [c \mu - [1 - \pi] \lambda(N - n_1 - n_2)]}, \\ \quad \left. \begin{array}{l} n_1 = N_1, \\ n_2 = 0, 1, \dots, N_2 - 1, \end{array} \right\} n_1 + n_2 > c, \\ \\ x_1(N_1 - 1, N_2) \pi \left[\frac{c \mu \tau_2 [1 - \tau_1] + \lambda(1) \tau_1}{c \mu \tau_2 - \pi c \mu \tau_2 [1 - \tau_1]} \right], \quad \begin{array}{l} n_1 = N_1, \\ n_2 = N_2, \end{array} \end{array} \right.$$

and

$$x_2(0, n_2) =$$

$$\left\{ \begin{array}{ll} [1-\pi] \left[\frac{\lambda(N)\tau_1}{n_2\mu[\tau_1+\tau_2-\tau_1\tau_2]} \right], & n_2 = 1, \\ x_2(0, n_2-1) [1-\pi] \left[\frac{[n_2-1]\mu\tau_2[1-\tau_1] + \lambda(N-n_2+1)\tau_1}{n_2\mu[\tau_1+\tau_2-\tau_1\tau_2]} \right] & n_2 = 2, 3, \dots, c-1, \\ x_2(0, n_2-1) [1-\pi] \tau_2 \left[\frac{[n_2-1]\mu\tau_2[1-\tau_1] + \lambda(N-n_2+1)\tau_1}{n_2\mu\tau_2 + \lambda(N-n_2)\tau_1[1-\tau_2]} \right], & n_2 = c, \\ x_2(0, n_2-1) [1-\pi] \left[\frac{c\mu\tau_2[1-\tau_1] + \lambda(N-n_2+1)\tau_1}{c\mu\tau_2 + \lambda(N-n_2)\tau_1[1-\tau_2]} \right], & n_2 = c+1, c+2, \dots, N_2-1, \\ x_2(0, n_2-1) [1-\pi] \left[\frac{c\mu\tau_2[1-\tau_1] + \lambda(N-n_2+1)\tau_1}{c\mu\tau_2 + \lambda(N-n_2)\tau_1[1-\tau_2]} \right] & \\ \frac{c\mu\tau_2 + \lambda(N-n_2)\tau_1[1-\tau_2] - \pi[c\mu\tau_2[1-\tau_1] + \lambda(N-n_2)\tau_1]}{\tau_1\tau_2[c\mu - \pi\lambda(N-n_2)]}, & n_2 = N_2. \end{array} \right. \quad [7.21]$$

Thus it follows that

$$\sum_{n_1=0}^{N_1} \sum_{n_2=0}^{N_2} \binom{n_1+n_2}{n_1} x_1(n_1, n_2) x_2(0, n_2) = \frac{\sum_{n_1=0}^{N_1} \sum_{n_2=0}^{N_2} p_N(n_1, n_2)}{p_N(0,0)} = \frac{1}{p_N(0,0)} - 1. \quad [7.22]$$

Before the analysis on the optimal flow control of a GE/GE/c finite capacity queue is presented, let us investigate the case when the

maximum number of packets is less than the number of servers (i.e. $c \geq N$). This is the GE/GE/ ∞ queue with ample (i.e. infinite) number of servers and no waiting buffer.

The Optimal Flow Control of a GE/GE/ ∞ Finite Capacity Queue

Suppose that centre 0 of Figure 78 is a GE/GE/ ∞ queue with finite number of packets. It follows that from relations [7.20] and [7.21], $x_1(n_1, n_2)$ and $x_2(0, n_2)$ are clearly given by

$$x_1(n_1, n_2) = \begin{cases} \pi \left[\frac{\lambda(N)\tau_1}{n_1 \mu [\tau_1 + \tau_2 - \tau_1 \tau_2]} \right], & n_1 = 1, n_2 = 0, \\ x_1(n_1 - 1, n_2) \pi \left[\frac{[n_1 + n_2 - 1] \mu \tau_2 (1 - \tau_1) + \lambda(N - n_1 - n_2 + 1) \tau_1}{[n_1 + n_2] \mu [\tau_1 + \tau_2 - \tau_1 \tau_2]} \right], & n_1 = 1, 2, \dots, N_1 - 1, \\ & n_2 = 0, 1, \dots, N_2 - 2, \\ x_1(n_1 - 1, n_2) \pi \left[\frac{[n_1 + n_2 - 1] \mu \tau_2 [1 - \tau_1] + \lambda(N - n_1 - n_2 + 1) \tau_1}{[n_1 + n_2] \mu \tau_2 + \lambda(N - n_1 - n_2) \tau_1 [1 - \tau_2]} \right] \\ \left[\frac{[n_1 + n_2] \mu \tau_2 + \lambda(N - n_1 - n_2) \tau_1 [1 - \tau_2]}{\tau_1 \tau_2 [n_1 + n_2] \mu - [1 - \pi] \lambda(N - n_1 - n_2)} \right] - \\ \left[\frac{(1 - \pi) [n_1 + n_2] \mu \tau_2 [1 - \tau_1] + \lambda(N - n_1 - n_2) \tau_1}{\tau_1 \tau_2 [n_1 + n_2] \mu - [1 - \pi] \lambda(N - n_1 - n_2)} \right] \right], & n_1 = N_1, \\ & n_2 = 0, 1, \dots, N_2 - 1, \\ x_1(N_1 - 1, N_2) \pi \left[\frac{[n_1 + n_2 - 1] \mu \tau_2 [1 - \tau_1] + \lambda(N - n_1 - n_2 + 1) \tau_1}{[n_1 + n_2] \mu \tau_2 - \pi [n_1 + n_2] \mu \tau_2 [1 - \tau_1]} \right], & n_1 = N_1, \\ & n_2 = N_2, \end{cases} \quad [7.23]$$

and

$$\begin{aligned}
 x_2(0, n_2) = & \\
 & \left\{ \begin{aligned}
 & [1-\pi] \left[\frac{\lambda(N)\tau_1}{n_2\mu[\tau_1+\tau_2-\tau_1\tau_2]} \right], & n_2 = 1, \\
 & x_2(0, n_2-1) [1-\pi] \left[\frac{[n_2-1]\mu\tau_2[1-\tau_1] + \lambda(N-n_2+1)\tau_1}{n_2\mu[\tau_1+\tau_2-\tau_1\tau_2]} \right] \\
 & \quad n_2 = 2, 3, \dots, N_2-1, \\
 & x_2(0, n_2-1) [1-\pi] \left[\frac{[n_2-1]\mu\tau_2[1-\tau_1] + \lambda(N-n_2+1)\tau_1}{n_2\mu\tau_2 + \lambda(N-n_2)\tau_1[1-\tau_2]} \right] \\
 & \quad \frac{n_2\mu\tau_2 + \lambda(N-n_2)\tau_1[1-\tau_2] - \pi[n_2\mu\tau_2[1-\tau_1] + \lambda(N-n_2)\tau_1]}{\tau_1\tau_2[n_2\mu - \pi\lambda(N-n_2)]}, & n_2 = N_2.
 \end{aligned} \right. \quad [7.24]
 \end{aligned}$$

Note that in this case no packets has to wait, the mean time delay equals to the mean service time $1/\mu$ and thus the control has no effect on the average time delay. This implies that in maximising the mean throughput, the time delay does not appear as an explicit parameter.

Theorem 7.1

The optimal control $\lambda_1 = \lambda_2(n_1)$, $n_1 = 1, 2, \dots, N_1$ and $\lambda_2 = \lambda_2(n_2)$, $n_2 = 1, 2, \dots, N_2$ of a GE/GE/ ∞ finite capacity queue at equilibrium with a maximum of N packets is given by $\lambda_1 = c_1$ and $\lambda_2 = c_2$ for all n_2 and n_2 . The corresponding maximum throughput is determine by

$$[\langle \chi \rangle N]_{\max} = \frac{\sum_{n_1=0}^{N_1} \sum_{n_2=0}^{N_2} [n_1+n_2] \mu \binom{n_1+n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)}{1 + \sum_{n_1=0}^{N_1} \sum_{n_2=0}^{N_2} \binom{n_1+n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)} \quad [7.25]$$

(n₁+n₂>0)

where

$$\hat{x}_1(n_1, n_2) = \left\{ \begin{array}{ll} \pi \left[\frac{\lambda_{\max} \tau_1}{n_1 \mu [\tau_1 + \tau_2 - \tau_1 \tau_2]} \right], & n_1 = 1, n_2 = 0, \\ \hat{x}_1(n_1-1, n_2) \pi \left[\frac{[n_1+n_2-1] \mu \tau_2 (1-\tau_1) + \lambda_{\max} \tau_1}{[n_1+n_2] \mu [\tau_1 + \tau_2 - \tau_1 \tau_2]} \right], & n_1 = 1, 2, \dots, N_1-1, \\ & n_2 = 0, 1, \dots, N_2-2, \\ \hat{x}_1(n_1-1, n_2) \pi \left[\frac{[n_1+n_2-1] \mu \tau_2 [1-\tau_1] + \lambda_{\max} \tau_1}{[n_1+n_2] \mu \tau_2 + \lambda_{\max} \tau_1 [1-\tau_2]} \right] \\ \left[\frac{[n_1+n_2] \mu \tau_2 + \lambda_{\max} \tau_1 [1-\tau_2] - (1-\pi) [n_1+n_2] \mu \tau_2 [1-\tau_1] + \lambda_{\max} \tau_1}{\tau_1 \tau_2 [n_1+n_2] \mu - [1-\pi] \lambda_{\max}} \right] & n_1 = N_1, \\ & n_2 = 0, 1, \dots, N_2-1, \\ \hat{x}_1(N_1-1, N_2) \pi \left[\frac{[n_1+n_2-1] \mu \tau_2 [1-\tau_1] + \lambda_{\max} \tau_1}{[n_1+n_2] \mu \tau_2 - \pi [n_1+n_2] \mu \tau_2 [1-\tau_1]} \right], & n_1 = N_1, \\ & n_2 = N_2, \end{array} \right. \quad [7.26]$$

and

$$\hat{x}_2(0, n_2) = \begin{cases} [1-\pi] \left[\frac{\lambda_{\max} \tau_1}{n_2 \mu [\tau_1 + \tau_2 - \tau_1 \tau_2]} \right], & n_2 = 1, \\ \hat{x}_2(0, n_2 - 1) [1-\pi] \left[\frac{[n_2 - 1] \mu \tau_2 [1 - \tau_1] + \lambda_{\max} \tau_1}{n_2 \mu [\tau_1 + \tau_2 - \tau_1 \tau_2]} \right] & n_2 = 2, 3, \dots, N_2 - 1, \\ x_2(0, n_2 - 1) [1-\pi] \left[\frac{[n_2 - 1] \mu \tau_2 [1 - \tau_1] + \lambda_{\max} \tau_1}{n_2 \mu \tau_2 + \lambda_{\max} \tau_1 [1 - \tau_2]} \right] & \\ \frac{n_2 \mu \tau_2 + \lambda_{\max} \tau_1 [1 - \tau_2] - \pi [n_2 \mu \tau_2 [1 - \tau_1] + \lambda_{\max} \tau_1]}{\tau_1 \tau_2 [n_2 \mu - \pi \lambda_{\max}]}, & n_2 = N_2. \end{cases} \quad [7.27]$$

Proof

Expression [7.25] is directly established by using formula [7.1] and [7.22]. It is sufficient to show that

$$\frac{\sum_{n_1=0}^{N_1} \sum_{n_2=0}^{N_2} [n_1 + n_2] \binom{n_1 + n_2}{n_1} x_1(n_1, n_2) x_2(0, n_2)}{\sum_{n_1=0}^{N_1} \sum_{n_2=0}^{N_2} \binom{n_1 + n_2}{n_1} x_1(n_1, n_2) x_2(0, n_2)} \leq 1$$

$$\frac{\sum_{n_1=0}^{N_1} \sum_{n_2=0}^{N_2} [n_1 + n_2] \binom{n_1 + n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)}{\sum_{n_1=0}^{N_1} \sum_{n_2=0}^{N_2} \binom{n_1 + n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)} \leq 1 \quad [7.28]$$

This can be written as

$$\begin{aligned}
 & \sum_{n_1=0}^{N_1} \sum_{n_2=0}^{N_2} [n_1+n_2] \binom{n_1+n_2}{n_1} \left[x_1(n_1, n_2) x_2(0, n_2) - \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2) \right] + \\
 & \sum_{n_1=0}^{N_1} \sum_{n_2=0}^{N_2} \sum_{m_1=0}^{N_1} \sum_{m_2=0}^{N_2} \binom{n_1+n_2}{n_1} \binom{m_1+m_2}{m_1} x_1(n_1, n_2) x_2(0, n_2) \hat{x}_1(m_1, m_2) \hat{x}_2(0, m_2) [n-m] \\
 & (n_1+n_2 > 0) \quad (m_1+m_2 > 0) \\
 & \leq 0.
 \end{aligned} \tag{7.29}$$

The latter inequality is true since the first term on the left hand side is negative in which $x_1(n_1, n_2) x_2(0, n_2) \leq \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)$ for $n_1+n_2 = 1, 2, \dots, N_1+N_2$ and the second term is also negative since its expression is symmetric in n and m and it can easily be shown that

$$\frac{x_1(n_1, n_2) x_2(0, n_2)}{x_1(m_1, m_2) x_2(0, m_2)} \leq \frac{\hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)}{\hat{x}_1(m_1, m_2) \hat{x}_2(0, m_2)} \tag{7.30}$$

is true for all $n > m$.

Corollary 7.1

The maximum throughput is strictly increasing with the number of packets N in the system, that is

$$[\langle \chi \rangle N]_{\max} < [\langle \chi \rangle N+1]_{\max} \tag{7.31}$$

for any integer N .

Proof

By using expression [7.25] it is sufficient to show that

$$\begin{aligned}
 & \frac{\sum_{n_1=0}^{N_1} \sum_{n_2=0}^{N_2} [n_1+n_2] \binom{n_1+n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)}{\sum_{n_1=0}^{N_1} \sum_{n_2=0}^{N_2} \binom{n_1+n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)} \leq \\
 & \frac{\sum_{n_1=0}^{N_1+1} \sum_{n_2=0}^{N_2} [n_1+n_2] \binom{n_1+n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)}{\sum_{n_1=0}^{N_1+1} \sum_{n_2=0}^{N_2} \binom{n_1+n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)} \quad [7.32]
 \end{aligned}$$

This can be written as

$$\begin{aligned}
 & \sum_{n_2=0}^{N_2} [N_1+1+n_2] \binom{N_1+n_2+1}{n_2} \hat{x}_1(N_1+1, n_2) \hat{x}_2(0, n_2) + \\
 & \sum_{i=0}^{N_2} \sum_{n_1=0}^{N_1} \sum_{n_2=0}^{N_2} \left[[N_1+1+i-n_1-n_2] \binom{N_1+1+i}{i} \binom{n_1+n_2}{n_1} \right. \\
 & \quad \left. \hat{x}_1(N_1+1, i) \hat{x}_2(0, i) \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2) \right] \geq 0. \quad [7.33]
 \end{aligned}$$

The latter inequality is true since $n_1+n_2 < N_1+N_2$.

The Optimal Flow Control of a GE/GE/c Finite Capacity Queue

Centre 0 of Figure 78 can be seen as a GE(n)/GE/c/N queueing system if the maximum number of packets is larger than the number of servers (i.e. $c < N$). For exposition purposes, each server is specified by $\mu(n) = \mu$, for all $n = 1, 2, \dots, N$ and constant coefficient of variation C_μ . In the case of $c \geq N$, this system is equivalent to GE(n)/GE/ ∞ /N queueing system presented in the previous section. Some results will be first presented in order to facilitate the proof of the main theorem on the optimal control that follows.

The maximum throughput of a GE/GE/c/N queueing system under a bounded time delay criterion is determined first. In the context of multiserver with finite capacity, the following lemmas are presented in order to facilitate the derivation of the main result.

Lemma 7.1

The control parameters $\lambda_1 = \lambda_1(n_1)$, $\lambda_2 = \lambda_2(n_2)$, $n_1 = 1, 2, \dots, N_1$, $n_2 = 1, 2, \dots, N_2$, are the optimum in the set of admissible controls for a given time delay T if they achieved the maximum of

$$\max \left[\frac{\sum_{n_1=0}^{N_1} \sum_{n_2=0}^{N_2} \mu(n_1+n_2) \binom{n_1+n_2}{n_1} x_1(n_1, n_2) x_2(0, n_2)}{1 + \sum_{n_1=0}^{N_1} \sum_{n_2=0}^{N_2} \binom{n_1+n_2}{n_1} x_1(n_1, n_2) x_2(0, n_2)} \right] \quad [7.34]$$

$(n_1+n_2 > 0)$
 $0 < n_1+n_2 \leq c$

subject to

$$\begin{aligned}
& \left[1 - \mu T \right] \sum_{n_1=0}^c \sum_{n_2=0}^c [n_1+n_2] \binom{n_1+n_2}{n_1} x_1(n_1, n_2) x_2(0, n_2) + \\
& \quad (n_1+n_2 > 0) \\
& \sum_{n_1=0}^{N_1} \sum_{n_2=0}^{N_2} \left[n_1+n_2 - c\mu T \right] \binom{n_1+n_2}{n_1} x_1(n_1, n_2) x_2(0, n_2) \leq 0 \quad [7.35] \\
& \quad (n_1+n_2 > c)
\end{aligned}$$

where

$$\begin{aligned}
0 \leq x_1(n_1, n_2) \leq \hat{x}_1(n_1, n_2) \quad n_1 = 1, 2, \dots, N_1, \\
n_2 = 0, 2, \dots, N_2, \quad [7.36]
\end{aligned}$$

and

$$0 \leq x_2(0, n_2) \leq \hat{x}_2(0, n_2) \quad n_2 = 1, 2, \dots, N_2, \quad [7.37]$$

and $\hat{x}_1(n_1, n_2)$ and $\hat{x}_2(0, n_2)$ are given by

$$\hat{x}_1(n_1, n_2) =$$

$$\left\{ \begin{array}{l} \pi \left[\frac{\lambda_{\max} \tau_1}{n_1 \mu [\tau_1 + \tau_2 - \tau_1 \tau_2]} \right], \quad n_1 = 1, \quad n_2 = 0, \\ \\ \hat{x}_1(n_1 - 1, n_2) \pi \left[\frac{[n_1 + n_2 - 1] \mu \tau_2 (1 - \tau_1) + \lambda_{\max} \tau_1}{[n_1 + n_2] \mu [\tau_1 + \tau_2 - \tau_1 \tau_2]} \right], \\ \quad \left. \begin{array}{l} n_1 = 1, 2, \dots, c-1, \\ n_2 = 0, 1, \dots, c-2, \end{array} \right\} 1 < n_1 + n_2 < c, \\ \\ \hat{x}_1(n_1 - 1, n_2) \pi \tau_2 \left[\frac{[c-1] \mu \tau_2 [1 - \tau_1] + \lambda_{\max} \tau_1}{c \mu \tau_2 + \lambda_{\max} \tau_1 [1 - \tau_2]} \right], \\ \quad \left. \begin{array}{l} n_1 = 1, 2, \dots, c, \\ n_2 = 0, 1, \dots, c-1, \end{array} \right\} n_1 + n_2 = c, \\ \\ \hat{x}_1(n_1 - 1, n_2) \pi \left[\frac{c \mu \tau_2 [1 - \tau_1] + \lambda_{\max} \tau_1}{c \mu \tau_2 + \lambda_{\max} \tau_1 [1 - \tau_2]} \right], \quad [7.38] \\ \quad \left. \begin{array}{l} n_1 = 1, 2, \dots, N_1 - 1, \\ n_2 = 0, 1, \dots, N_2, \end{array} \right\} n_1 + n_2 > c, \\ \\ \hat{x}_1(n_1 - 1, n_2) \pi \left[\frac{c \mu \tau_2 [1 - \tau_1] + \lambda_{\max} \tau_1}{c \mu \tau_2 + \lambda_{\max} \tau_1 [1 - \tau_2]} \right], \\ \frac{c \mu \tau_2 + \lambda_{\max} \tau_1 [1 - \tau_2] - (1 - \pi) [c \mu \tau_2 [1 - \tau_1] + \lambda_{\max} \tau_1]}{\tau_1 \tau_2 [c \mu - [1 - \pi] \lambda_{\max}]}, \\ \quad \left. \begin{array}{l} n_1 = N_1, \\ n_2 = 0, 1, \dots, N_2 - 1, \end{array} \right\} n_1 + n_2 > c, \\ \\ \hat{x}_1(N_1 - 1, N_2) \pi \left[\frac{c \mu \tau_2 [1 - \tau_1] + \lambda_{\max} \tau_1}{c \mu \tau_2 - \pi c \mu \tau_2 [1 - \tau_1]} \right], \quad \begin{array}{l} n_1 = N_1, \\ n_2 = N_2, \end{array} \end{array} \right.$$

and

$$\hat{x}_2(0, n_2) =$$

$$\left\{ \begin{array}{l} [1-\pi] \left[\frac{\lambda_{\max} \tau_1}{n_2 \mu [\tau_1 + \tau_2 - \tau_1 \tau_2]} \right], \quad n_2 = 1, \\ \hat{x}_2(0, n_2 - 1) [1-\pi] \left[\frac{[n_2 - 1] \mu \tau_2 [1 - \tau_1] + \lambda_{\max} \tau_1}{n_2 \mu [\tau_1 + \tau_2 - \tau_1 \tau_2]} \right], \quad n_2 = 2, 3, \dots, c-1, \\ \hat{x}_2(0, n_2 - 1) [1-\pi] \tau_2 \left[\frac{[n_2 - 1] \mu \tau_2 [1 - \tau_1] + \lambda_{\max} \tau_1}{n_2 \mu \tau_2 + \lambda_{\max} \tau_1 [1 - \tau_2]} \right], \quad n_2 = c, \\ \hat{x}_2(0, n_2 - 1) [1-\pi] \left[\frac{c \mu \tau_2 [1 - \tau_1] + \lambda_{\max} \tau_1}{c \mu \tau_2 + \lambda_{\max} \tau_1 [1 - \tau_2]} \right], \quad n_2 = c+1, c+2, \dots, N_2 - 1, \\ \hat{x}_2(0, n_2 - 1) [1-\pi] \left[\frac{c \mu \tau_2 [1 - \tau_1] + \lambda_{\max} \tau_1}{c \mu \tau_2 + \lambda_{\max} \tau_1 [1 - \tau_2]} \right], \quad n_2 = N_2. \end{array} \right. \quad [7.39]$$

Proof

From equation [7.2] and relation [7.22] we have

$$\langle \tau \rangle N = \frac{\sum_{n_1=0}^{N_1} \sum_{n_2=0}^{N_2} [n_1 + n_2] p_N(n_1, n_2)}{\sum_{n_1=0}^{N_1} \sum_{n_2=0}^{N_2} \mu(n_1 + n_2) p_N(n_1, n_2)} =$$

$$\begin{aligned}
& \frac{\sum_{n_1=0}^{N_1} \sum_{n_2=0}^{N_2} \binom{n_1+n_2}{n_1} [n_1+n_2] x_1(n_1, n_2) x_2(0, n_2)}{(n_1+n_2 > 0)} \leq T \quad [7.40] \\
& \frac{\sum_{n_1=0}^{N_1} \sum_{n_2=0}^{N_2} \mu(n_1+n_2) \binom{n_1+n_2}{n_1} x_1(n_1, n_2) x_2(0, n_2)}{(n_1+n_2 > 0)}
\end{aligned}$$

or equivalently, inequality [7.35] holds and from [7.1] and [7.22], it follows that the relation [7.34] is satisfied. A bounded expression for a maximum time delay is determined as follows.

Lemma 7.2

The maximum mean time delay $T_{\max}(L_1, L_2)$ achieved with L_1 and L_2 of class 1 and 2 packets in a $GE(n)/GE/c/N$ queueing system ($c < L_1 + L_2 \leq N_1 + N_2$) is given by

$$T_{\max}(L_1, L_2) = \frac{\sum_{n_1=0}^{L_1} \sum_{n_2=0}^{L_2} \binom{n_1+n_2}{n_1} [n_1+n_2] \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)}{\hat{z}(L_1, L_2)} \leq \frac{L_1 + L_2}{c\mu}, \quad [7.41]$$

where $\hat{z}(L_1, L_2)$ is given by

$$\hat{z}(L_1, L_2) = \mu \left[\sum_{n_1=0}^c \sum_{n_2=0}^c [n_1+n_2] \binom{n_1+n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2) \right]_{0 < n_1+n_2 \leq c}$$

$$+ \sum_{\substack{n_1=0 \\ (n_1+n_2 > c)}}^{L_1} \sum_{n_2=0}^{L_2} c \binom{n_1+n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2) \quad [7.42]$$

Proof

Let us first show that

$$\frac{\sum_{n_1=0}^{L_1} \sum_{\substack{n_2=0 \\ (n_1+n_2 > 0)}}^{L_2} \binom{n_1+n_2}{n_1} [n_1+n_2] x_1(n_1, n_2) x_2(0, n_2)}{z(L_1, L_2)} \leq$$

$$\frac{\sum_{n_1=0}^{L_1} \sum_{\substack{n_2=0 \\ (n_1+n_2 > 0)}}^{L_2} \binom{n_1+n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)}{\hat{z}(L_1, L_2)}, \quad [7.43]$$

where $z(L_1, L_2)$ is given by substituting $\hat{x}_1(n_1, n_2)$ and $\hat{x}_2(0, n_2)$ by $x_1(n_1, n_2)$ and $x_2(0, n_2)$ respectively. This can be written as

$$\sum_{\substack{n_1=0 \\ 0 < n_1+n_2 \leq c}}^{L_1} \sum_{\substack{n_2=0 \\ m_1+m_2 > c}}^{L_2} \sum_{m_1=0}^{L_1} \sum_{m_2=0}^{L_2} \left[[n_1+n_2] [c-m_1-m_2] \binom{n_1+n_2}{n_1} \binom{m_1+m_2}{m_1} \right]$$

$$x_1(n_1, n_2) x_2(0, n_2) \hat{x}_1(m_1, m_2) \hat{x}_2(0, m_2) - \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2) x_1(m_1, m_2) x_2(0, m_2) \Big]$$

$$+ c \sum_{n_1=0}^{L_1} \sum_{\substack{n_2=0 \\ n_1+n_2 > c}}^{L_2} \sum_{m_1=0}^{L_1} \sum_{\substack{m_2=0 \\ m_1+m_2 > c}}^{L_2} \left[[n_1+n_2-m_1-m_2] \binom{n_1+n_2}{n_1} \binom{m_1+m_2}{m_1} \right]$$

$$\left[x_1(n_1, n_2) x_2(0, n_2) \hat{x}_1(m_1, m_2) \hat{x}_2(0, m_2) \right] \leq 0. \quad [7.44]$$

The first term on the left hand side of inequality [7.44] is negative

since $x_1(n_1, n_2) x_2(0, n_2) \hat{x}_1(m_1, m_2) \hat{x}_2(0, m_2) \leq \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2) x_1(m_1, m_2) x_2(0, m_2)$ for all $c < m_1 + m_2 \leq L_1 + L_2$ and the second term is also negative since expression is symmetric n_1, n_2, m_1 and m_2 and the fact that

$$\frac{x_1(n_1, n_2) x_2(0, n_2)}{x_1(m_1, m_2) x_2(0, m_2)} \leq \frac{\hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)}{\hat{x}_1(m_1, m_2) \hat{x}_2(0, m_2)} \quad [7.45]$$

for all $n > m$ and therefore the left hand side of expression [7.44] is negative. Note also that

$$T_{\max}(L_1, L_2) \leq \frac{L_1 + L_2}{c\mu}. \quad [7.46]$$

Since inequality [7.41] can be rewritten as

$$\begin{aligned} & [L_1 + L_2 - c] \sum_{\substack{n_1=0 \\ 0 < n_1 + n_2 \leq c}}^c \sum_{n_2=0}^c [n_1 + n_2] \binom{n_1 + n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2) + \\ & c \sum_{\substack{n_1=0 \\ (n_1 + n_2 > c)}}^{L_1} \sum_{n_2=0}^{L_2} \left[[L_1 + L_2] - [n_1 + n_2] \right] \binom{n_1 + n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2) \geq 0, \end{aligned} \quad [7.47]$$

The latter inequality is true since $n_1 + n_2 \leq L_1 + L_2$. Note that the bound of $T_{\max}(L_1, L_2)$ applies to the case when the packets after receiving

service instantaneously rejoin the queue. Then $(L_1+L_2)/c\mu$ is exactly the mean time delay of a $GE(n)/GE/c/N$ queue constantly having L_1+L_2 packets in the system. Moreover, the following lemma holds.

Lemma 7.3

The maximum mean time delay $T_{\max}(L_1, L_2)$ in the $GE(n)/GE/c/N$ system is strictly increasing with the number of packets L_1 and L_2 i.e.

$$T_{\max}(L_1, L_2) < T_{\max}(L_1+1, L_2) \quad [7.48]$$

for all L_1 and L_2 , $L_1+L_2 = c+1, c+2, \dots, N$.

Proof

It is sufficient to show that

$$\frac{\sum_{n_1=0}^{L_1} \sum_{n_2=0}^{L_2} \binom{n_1+n_2}{n_1} [n_1+n_2] \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)}{\hat{z}(L_1, L_2)} < \frac{\sum_{n_1=0}^{L_1+1} \sum_{n_2=0}^{L_2} \binom{n_1+n_2}{n_1} [n_1+n_2] \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)}{\hat{z}(L_1+1, L_2)}, \quad [7.49]$$

which can be written as

$$\sum_{n_1=0}^c \sum_{n_2=0}^c \left[[n_1+n_2] \binom{n_1+n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2) \right] +$$

$$0 < n_1 + n_2 \leq c$$

$$\sum_{n_2=0}^{L_2} \left[\binom{L_1+1+n_2}{n_2} [L_1+1+n_2-c] \hat{x}_1(L_1+1, n_2) \hat{x}_2(0, n_2) \right] +$$

$$\sum_{n_1=0}^{L_1} \sum_{n_2=0}^{L_2} \sum_{i=0}^{L_2} \left[\binom{L_1+1+i}{i} \hat{x}_1(L_1+1, i) \hat{x}_2(0, i) \binom{n_1+n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2) \right]$$

$$(n_1 + n_2 > c)$$

$$\left[[L_1+1+i-n_1-n_2] \right] > 0. \quad [7.50]$$

This inequality is true since $n_1 < L_1+1$. In this context, the following theorem is presented.

Theorem 7.2

Given that $T_{\max}(L_1-1, L_2) < T \leq T_{\max}(L_1, L_2)$, $c+1 \leq L_1+L_2 \leq N_1+N_2$, the optimal control of a GE(n)/GE/c finite capacity queue with two classes of users and a maximum of $N=N_1+N_2$ controlled packets in the system are given by

$$\begin{aligned}
& \lambda(N-n_1-n_2+1) = \\
& \left\{ \begin{array}{l} \lambda_{\max}, \quad \left. \begin{array}{l} n_1 = 0, 1, \dots, L_1, \\ n_2 = 0, 1, \dots, L_2, \\ n_1+n_2 = 0, 1, \dots, L_1+L_2-1, \end{array} \right\} \\ \\ \left[\begin{array}{l} \sum_{i_1=0}^c \sum_{i_2=0}^c \frac{[\mu T-1][i_1+i_2] \binom{i_1+i_2}{i_1} \hat{x}_1(i_1, i_2) \hat{x}_2(0, i_2)}{[L_1+L_2-c\mu T] \binom{L_1+L_2}{L_1}} + \\ \sum_{\substack{j_1=0 \\ (j_1+j_2 > c)}}^{L_1-1} \sum_{j_2=0}^{L_2} \frac{[c\mu T-j_1-j_2] \binom{j_1+j_2}{j_1} \hat{x}_1(j_1, j_2) \hat{x}_2(0, j_2)}{[L_1+L_2-c\mu T] \binom{L_1+L_2}{L_1}} \hat{x}_1(L_1-1, L_2) \hat{x}_2(0, L_2) \end{array} \right] \\ \\ \left[\frac{c\mu\tau_2 - \pi c\mu\tau_2 [1-\tau_1]}{\tau_1} \right] - \frac{c\mu\tau_2 (1-\tau_1)}{\tau_1}, \quad \left. \begin{array}{l} n_1 = L_1, \\ n_2 = L_2, \\ n_1+n_2 = L_1+L_2, \end{array} \right\} \\ \\ \left. \begin{array}{l} n_1 = L_1, L_1+1, \dots, N_1, \\ n_2 = L_2, L_2+1, \dots, N_2, \\ n_1+n_2 = L_1+L_2+1, \dots, N_1+N_2, \end{array} \right\} \\ 0, \end{array} \right\} \quad [7.51]
\end{aligned}$$

and the maximum throughput is given by

$$F(T) = \frac{\hat{z}(L_1, L_2)}{1 + \sum_{\substack{n_1=0 \\ (n_1+n_2 > 0)}}^{L_1} \sum_{n_2=0}^{L_2} \binom{n_1+n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)} \quad [7.52]$$

where L_1+L_2 is the maximum allowed number of packets in the system (i.e. "sliding" window size).

Proof

Let us assume that

$$T_{\max}(L_1-1, L_2) < T \leq T_{\max}(L_1, L_2) \quad [7.53]$$

which is true as shown in Lemma 7.3 above for all $L_1+L_2 = c+1, c+2, \dots, N$. The condition for a bounded maximum mean time delay in [7.35] implies that

$$(1-\mu T) \sum_{n_1=0}^c \sum_{n_2=0}^c [n_1+n_2] \binom{n_1+n_2}{n_1} x_1(n_1, n_2) x_2(0, n_2) +$$

$$0 < n_1+n_2 \leq c$$

$$\sum_{n_1=0}^{N_1} \sum_{n_2=0}^{N_2} (n_1+n_2 - c\mu T) \binom{n_1+n_2}{n_1} x_1(n_1, n_2) x_2(0, n_2) +$$

$$\begin{matrix} (n_1+n_2 > c) \\ (n_1+n_2 \neq L_1+L_2) \end{matrix}$$

$$[L_1+L_2-c\mu T] \binom{L_1+L_2}{L_1} x_1(L_1, L_2) x_2(0, L_2) \leq 0, \quad [7.54]$$

which can be written as

$$x_1(L_1, L_2) x_2(0, L_2) \leq - \frac{\sum_{n_1=0}^{N_1} \sum_{n_2=0}^{N_2} [n_1+n_2 - c\mu T] \binom{n_1+n_2}{n_1} x_1(n_1, n_2) x_2(0, n_2)}{(L_1+L_2-c\mu T) \binom{L_1+L_2}{L_1}} -$$

$$\begin{matrix} (n_1+n_2 > c) \\ (n_1+n_2 \neq L_1+L_2) \end{matrix}$$

$$\begin{aligned}
 & x_1(n_1, n_2) x_2(0, n_2) = \\
 & \left\{ \begin{aligned}
 & 0, \quad \begin{aligned}
 & n_1 = L_1, \dots, N_1, \\
 & n_2 = L_2, \dots, N_2, \\
 & n_1 + n_2 > L_1 + L_2,
 \end{aligned} \\
 & \sum_{\substack{c \\ i_1=0 \\ 0 < i_1+i_2 \leq c}}^c \sum_{\substack{c \\ i_2=0}}^c \frac{[\mu T - 1][i_1 + i_2] \binom{i_1 + i_2}{i_1} \hat{x}_1(i_1, i_2) \hat{x}_2(0, i_2)}{(L_1 + L_2 - c\mu T) \binom{L_1 + L_2}{L_1}} + \\
 & \sum_{\substack{L_1-1 \\ j_1=0 \\ (j_1+j_2 > c)}}^{L_1-1} \sum_{\substack{L_2 \\ j_2=0}}^{L_2} \frac{[c\mu T - (j_1 + j_2)] \binom{j_1 + j_2}{j_1} \hat{x}_1(j_1, j_2) \hat{x}_2(0, j_2)}{[L_1 + L_2 - c\mu T] \binom{L_1 + L_2}{L_1}}, \\
 & \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2), \quad \begin{aligned}
 & n_1 = 0, 1, \dots, L_1, \\
 & n_2 = 0, 1, \dots, L_2, \\
 & 0 < n_1 + n_2 < L_1 + L_2,
 \end{aligned}
 \end{aligned} \right. \quad [7.57]
 \end{aligned}$$

It can be shown that

$$0 < x_1(L_1, L_2) x_2(0, L_2) \leq \hat{x}_1(L_1, L_2) \hat{x}_2(0, L_2) \quad [7.58]$$

can be achieved by substituting [7.57] into [7.58]. It follows that

$$0 < \sum_{\substack{c \\ n_1=0 \\ 0 < n_1+n_2 \leq c}}^c \sum_{\substack{c \\ n_2=0}}^c \frac{[\mu T - 1][n_1 + n_2] \binom{n_1 + n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)}{(L_1 + L_2 - c\mu T) \binom{L_1 + L_2}{L_1}} +$$

$$\sum_{\substack{n_1=0 \\ (n_1+n_2 > c)}}^{L_1-1} \sum_{n_2=0}^{L_2} \frac{[c\mu T - (n_1+n_2)] \binom{n_1+n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)}{[L_1+L_2 - c\mu T] \binom{L_1+L_2}{L_1}} \leq \hat{x}_1(L_1, L_2) \hat{x}_2(0, L_2) \quad [7.59]$$

which is equivalent to

$$\frac{\sum_{n_1=0}^{L_1-1} \sum_{n_2=0}^{L_2} \binom{n_1+n_2}{n_1} (n_1+n_2) \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)}{\hat{z}(L_1-1, L_2)} < T \leq$$

$$\frac{\sum_{n_1=0}^{L_1-1} \sum_{n_2=0}^{L_2} \binom{n_1+n_2}{n_1} (n_1+n_2) \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)}{\hat{z}(L_1-1, L_2) + c\mu \binom{L_1+L_2}{L_1} \hat{x}_1(L_1, L_2) \hat{x}_2(0, L_2)} +$$

$$\frac{(L_1+L_2) \binom{L_1+L_2}{L_1} \hat{x}_1(L_1, L_2) \hat{x}_2(0, L_2)}{\hat{z}(L_1-1, L_2) + c\mu \binom{L_1+L_2}{L_1} \hat{x}_1(L_1, L_2) \hat{x}_2(0, L_2)}, \quad [7.60]$$

or from Lemma 7.3, the above can be written as in [7.53]. Then it can be shown from [7.57] that Theorem 7.2 is true as follows. Since from [7.57]

$$x_1(1, 0) x_2(0, 0) = \hat{x}_1(1, 0) \hat{x}_2(0, 0) \quad [7.61]$$

or from [7.20], [7.21], [7.38] and [7.39]

$$\left[\frac{\pi \lambda(N) \tau_1}{n_1 \mu [\tau_1 + \tau_2 - \tau_1 \tau_2]} \right] \left[\frac{(1-\pi) \lambda(N) \tau_1}{n_1 \mu [\tau_1 + \tau_2 - \tau_1 \tau_2]} \right] =$$

$$\left[\frac{\pi \lambda_{\max} \tau_1}{n_1 \mu [\tau_1 + \tau_2 - \tau_1 \tau_2]} \right] \left[\frac{(1-\pi) \lambda_{\max} \tau_1}{n_1 \mu [\tau_1 + \tau_2 - \tau_1 \tau_2]} \right] \quad [7.62]$$

which can be shown that

$$\lambda(N) = \lambda_{\max} \quad [7.63]$$

It follows from [7.57] that

$$x_1(n_1, n_2) x_2(0, n_2) = \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2), \quad 2 \leq n_1 + n_2 \leq L_1 + L_2 - 1 \quad [7.64]$$

and from the above relationship, it can be similarly shown that [7.51] holds for $n_1 + n_2 = 2, 3, \dots, L_1 + L_2 - 1$. When the system has achieved maximum capacity, it is implied that

$$\begin{aligned} x_1(L_1, L_2) x_2(0, L_2) = & \sum_{\substack{n_1=0 \\ 0 < n_1 + n_2 \leq c}}^c \sum_{\substack{n_2=0 \\ 0 < n_1 + n_2 \leq c}}^c \frac{[\mu T - 1] \binom{n_1 + n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)}{(L_1 + L_2 - c \mu T) \binom{L_1 + L_2}{L_1}} + \\ & \sum_{\substack{n_1=0 \\ (n_1 + n_2 > c)}}^{L_1 - 1} \sum_{\substack{n_2=0 \\ (n_1 + n_2 > c)}}^{L_2} \frac{[c \mu T - (n_1 + n_2)] \binom{n_1 + n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)}{[L_1 + L_2 - c \mu T] \binom{L_1 + L_2}{L_1}}, \end{aligned} \quad [7.65]$$

which is equivalent to

which is equivalent to

$$x_1(L_1, L_2) x_2(0, L_2) = \hat{x}_1(L_1-1, L_2) \hat{x}_2(0, L_2) \pi \left[\frac{c\mu\tau_2(1-\tau_1) + \lambda(N-n_1+1-n_2)\tau_1}{c\mu\tau_2 - \pi c\mu\tau_2[1-\tau_1]} \right] \quad [7.66]$$

and thus it follows that [7.51] holds for $n_1+n_2 = L_1+L_2$. Finally the maximum throughput given by [7.52] is achieved if the equalities in [7.57] are satisfied. Note that, since the queueing centre 0 of figure 78 will contain at all times at least $N - L_1 - L_2$ packets, the optimum control can also be achieved by a control scheme using a total of $N = L_1+L_2$ packets (i.e. fixed window scheme). It is required that

$$\lambda(N-n_1-n_2+1) = \left\{ \begin{array}{l} \lambda_{\max}, \quad \left. \begin{array}{l} n_1 = 0, 1, \dots, L_1, \\ n_2 = 0, 1, \dots, L_2, \\ n_1+n_2 = 0, 1, \dots, L_1+L_2-1, \end{array} \right\} \\ \left[\begin{array}{l} \sum_{i_1=0}^c \sum_{i_2=0}^c \frac{[\mu T-1][i_1+i_2] \binom{i_1+i_2}{i_1} \hat{x}_1(i_1, i_2) \hat{x}_2(0, i_2)}{[L_1+L_2-c\mu T] \binom{L_1+L_2}{L_1} \hat{x}_1(L_1-1, L_2) \hat{x}_2(0, L_2)} + \\ \sum_{\substack{j_1=0 \\ (j_1+j_2 > c)}}^{L_1-1} \sum_{j_2=0}^{L_2} \frac{[c\mu T-j_1-j_2] \binom{j_1+j_2}{j_1} \hat{x}_1(j_1, j_2) \hat{x}_2(0, j_2)}{[L_1+L_2-c\mu T] \binom{L_1+L_2}{L_1} \hat{x}_1(L_1-1, L_2) \hat{x}_2(0, L_2)} \end{array} \right] \\ \left[\frac{c\mu\tau_2 - \pi c\mu\tau_2[1-\tau_1]}{\tau_1} \right] - \frac{c\mu\tau_2(1-\tau_1)}{\tau_1}, \quad \left. \begin{array}{l} n_1 = L_1, \\ n_2 = L_2, \\ n_1+n_2 = L_1+L_2, \end{array} \right\} \end{array} \right. \quad [7.67]$$

Therefore from [7.51] it can be seen that the optimal flow control of a GE(n)/GE/c finite capacity queue is a sliding (variable) window type mechanism which is known in control theory as a bang-bang control (i.e. a control with two possible active states), and the optimum offered load is as follows:

- i. equal to the maximum input rate $\lambda_{\max}$ when the number of packets is less than $L_1 + L_2$,
- ii. less than the maximum input rate when the number of packets is equal to $L_1 + L_2$, and
- iii. equal to zero when the system is full.

The window size can be derived from the maximum time delay of the system via the following corollary.

Corollary 7.2

If $T = T_{\max}(L)$, then the throughput time delay function and the maximum time delay are given by

$$F[T_{\max}(L_1, L_2)] = \frac{\sum_{n_1=0}^{L_1} \sum_{n_2=0}^{L_2} \mu(n_1+n_2) \binom{n_1+n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)}{1 + \sum_{n_1=0}^{L_1} \sum_{n_2=0}^{L_2} \binom{n_1+n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)}, \quad [7.68]$$

and

$$T_{\max}(L_1, L_2) = \frac{\sum_{n_1=0}^{L_1} \sum_{n_2=0}^{L_2} \binom{n_1+n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)}{\sum_{n_1=0}^{L_1} \sum_{n_2=0}^{L_2} \mu(n_1+n_2) \binom{n_1+n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)}, \quad [7.69]$$

Proof

For a given $T = T_{\max}(L_1, L_2)$, it follows that

$$\begin{aligned} x_1(L_1, L_2) x_2(0, L_2) &= \sum_{\substack{n_1=0 \\ 0 < n_1+n_2 \leq c}}^c \sum_{n_2=0}^c \frac{[\mu T - 1] \binom{n_1+n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)}{(L_1 + L_2 - c\mu T) \binom{L_1+L_2}{L_1}} + \\ &\quad \sum_{\substack{n_1=0 \\ (n_1+n_2 > c)}}^{L_1-1} \sum_{n_2=0}^{L_2} \frac{[c\mu T - (n_1+n_2)] \binom{n_1+n_2}{n_1} \hat{x}_1(n_1, n_2) \hat{x}_2(0, n_2)}{[L_1 + L_2 - c\mu T] \binom{L_1+L_2}{L_1}} \\ &= \hat{x}_1(L_1, L_2) \hat{x}_2(0, L_2) \end{aligned} \quad [7.70]$$

and from equations [7.1], [7.2], [7.5] and [7.22], relations [7.68] and [7.69] are obtained by satisfying equations [7.38] and [7.70].

Numerical Examples

In this section numerical examples are presented to illustrate how critically the throughput time delay function, $F[T_{\max}(L_1, L_2)]$ is

affected by the distributional form of the interarrival time and service time distributions and the maximum input rate. The GE distributional model is chosen to approximate general distributions with known first two moments.

The maximum throughput and the maximum time delay as a function of the control are depicted in Figures 80-85 for C_μ^2 , $C_{\lambda_1}^2$ and $C_{\lambda_2}^2$ varying from 0.5 to 2.0, $c_1 = 0.6$, $c_2 = 1.0$, $\mu = 1$ and $c = 2$. It is noted that the GE-type solutions for the maximum throughput as a function of the control provides lower bounds (pessimistic) when C_μ^2 , $C_{\lambda_1}^2$ and $C_{\lambda_2}^2 = 2$ and upper bounds (optimistic) when C_μ^2 , $C_{\lambda_1}^2$ and $C_{\lambda_2}^2 = 0.5$ (see Figures 80-82). As expected, the maximum time delay as a function of the control provides lower bounds (pessimistic) when C_μ^2 , $C_{\lambda_1}^2$ and $C_{\lambda_2}^2 = 0.5$ and upper bounds (optimistic) when C_μ^2 , $C_{\lambda_1}^2$ and $C_{\lambda_2}^2 = 2.0$ (see Figures 83-85).

In the second example, the maximum throughput and the maximum time delay as a function of the control are depicted in Figures 86-95 for C_μ^2 , $C_{\lambda_1}^2$ and $C_{\lambda_2}^2$ varying from 0.5 to 5.0, $c_1 = 1.2$, c_2 varying from 0.6 to 1.4, $\mu = 1$ and $c = 3$. It is noted that the GE-type solutions for the maximum throughput as a function of the control provides lower bounds (pessimistic) when C_μ^2 , $C_{\lambda_1}^2$ and $C_{\lambda_2}^2 = 5.0$ and upper bounds (optimistic) when C_μ^2 , $C_{\lambda_1}^2$ and $C_{\lambda_2}^2 = 0.5$ (see Figures 86-88). The maximum throughput as a function of the control for $c_1 + c_2$ varying from 1.8 to 2.6 with C_μ^2 , $C_{\lambda_1}^2$ and $C_{\lambda_2}^2 = 5.0$, $\mu = 1$, are depicted in Figures 88-90. It can be observed that the GE(n)/GE/c/N solution predicts the increasing optimistic values as expected for throughput as $c_1 + c_2$ moves between 1.8 to 2.6. Also as expected, the maximum time delay as a

function of the control provides lower bounds (pessimistic) when C_{μ}^2 , $C_{\lambda_1}^2$ and $C_{\lambda_2}^2 = 0.5$ and upper bounds (optimistic) when C_{μ}^2 , $C_{\lambda_1}^2$ and $C_{\lambda_2}^2 = 5.0$ (see Figures 91-93). It can also be observed that the GE(n)/GE/c/N solution predicts the increasing optimistic values for maximum time delay as $c_1 + c_2$ moves between 1.8 to 2.6 (see Figures 93-95).

The relationship between the maximum time delay and the maximum throughput can be shown through their dependence on the control. To do so, the control of user class 1 is parameterised and the control of user class 2 is increased. Then, the control of user class 2 is parameterised and that of user class 1 is increased. The resulting throughput-time delay trade-off are shown in Figures 96 and 97 respectively for C_{μ}^2 , $C_{\lambda_1}^2$ and $C_{\lambda_2}^2 = 1.0$, $\mu = 1$, $c = 3$. The optimal control is obtained by comparing these curves and choosing the control which corresponds to a maximum throughput for a given time delay constraints. By superimposing the curves in Figures 96 and 97, it can be seen that the optimal control allows near window sizes for the two users. If $c_1 < c_2$, one user class 2 packet should enter the system first. If the time delay constraint has not yet been achieved, one user class 1 packet has to enter the system. This process is repeated alternating between user class 1 and 2 until the time delay constraint is achieved. Again as stated in Chapter VI, it is important to note that only the order relation between c_1 and c_2 is needed in the determination of the optimal window sizes for the two users.

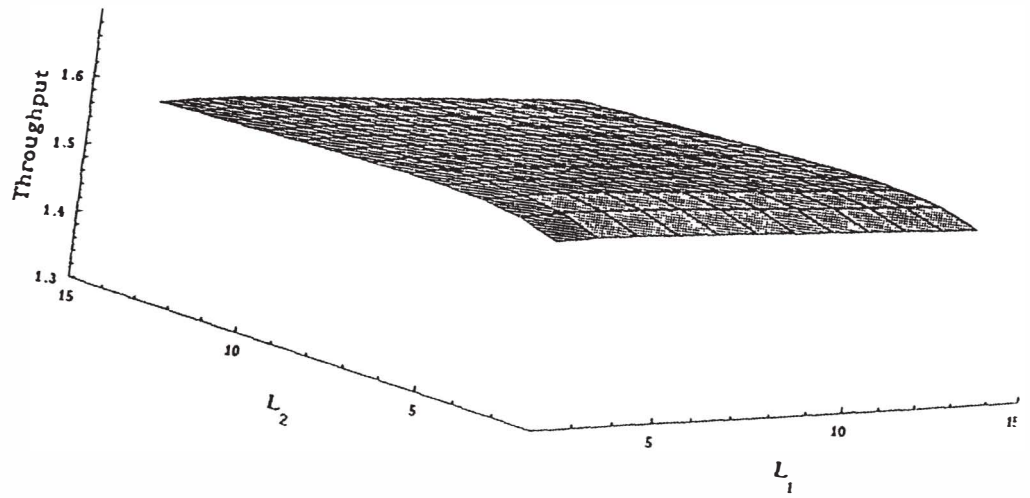


Figure 80 : Throughput versus Control
 $(C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 0.5, c_1 = 0.6, c_2 = 1.0, \mu = 1, c = 2)$

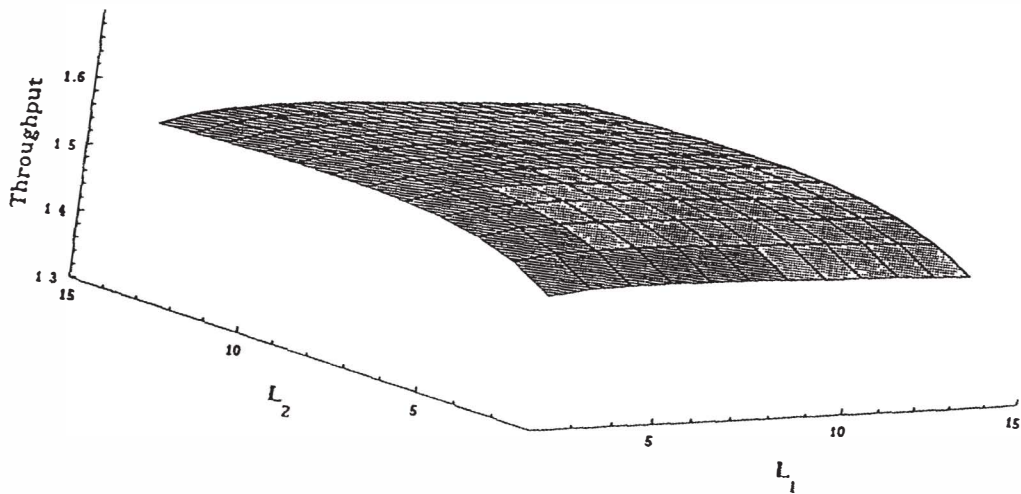


Figure 81 : Throughput versus Control
 $(C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 1.0, c_1 = 0.6, c_2 = 1.0, \mu = 1, c = 2)$

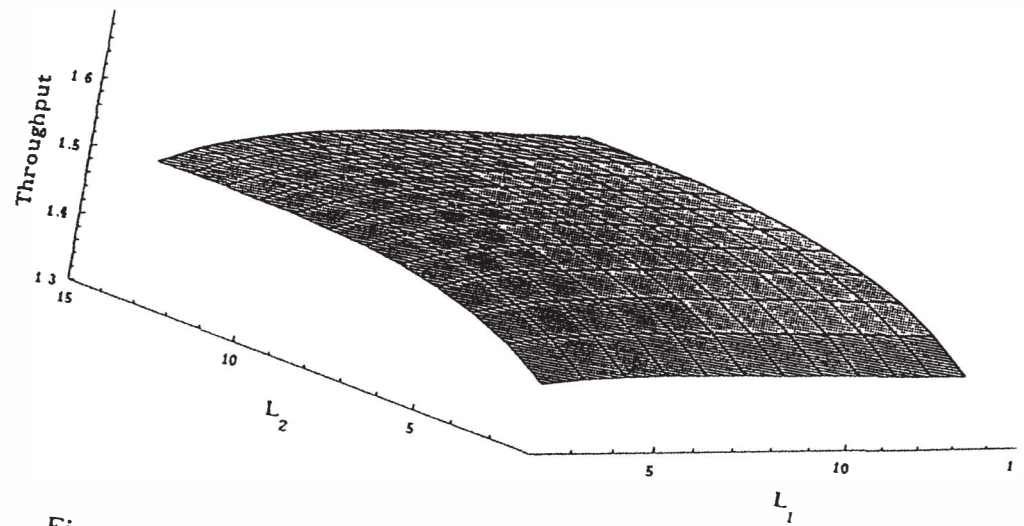


Figure 82 : Throughput versus Control
 $(C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 2.0, c_1 = 0.6, c_2 = 1.0, \mu = 1, c = 2)$

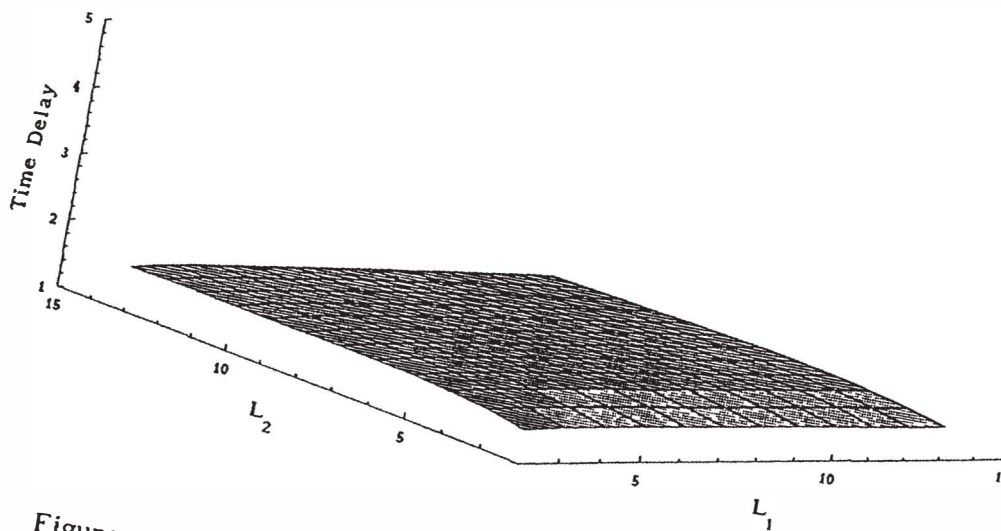


Figure 83 : Time Delay versus Control
 $(C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 0.5, c_1 = 0.6, c_2 = 1.0, \mu = 1, c = 2)$

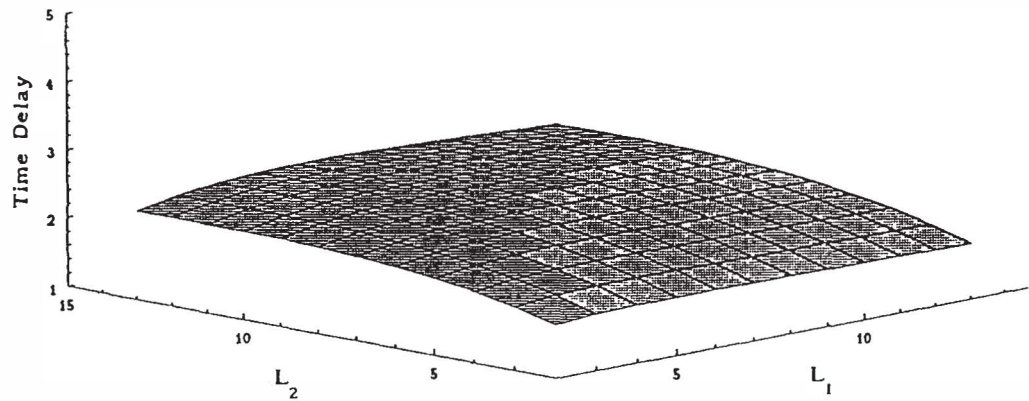


Figure 84 : Time Delay versus Control

$$(C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 1.0, c_1 = 0.6, c_2 = 1.0, \mu = 1, c = 2)$$

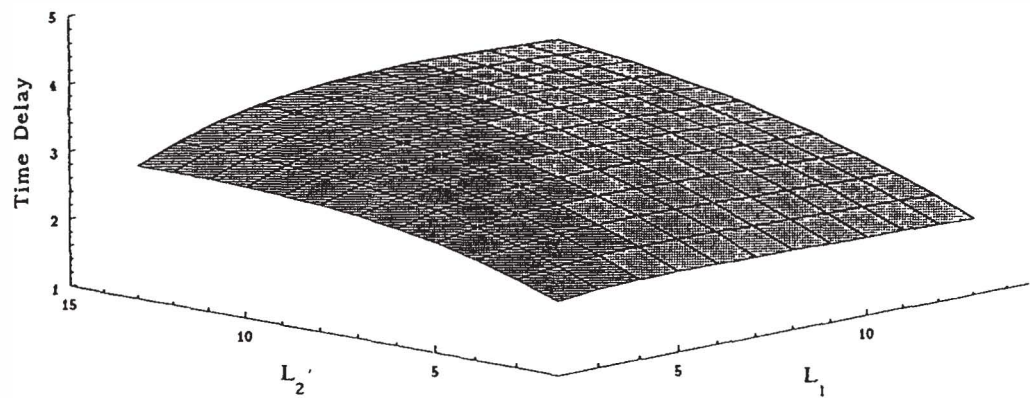


Figure 85 : Time Delay versus Control

$$(C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 2.0, c_1 = 0.6, c_2 = 1.0, \mu = 1, c = 2)$$

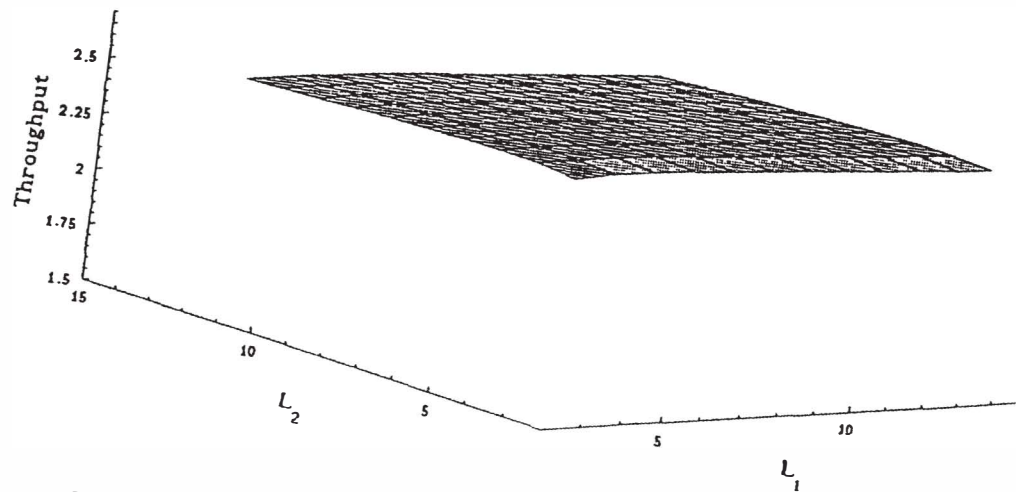


Figure 86 : Throughput versus Control
 $(C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 0.5, c_1 = 1.2, c_2 = 1.4, \mu = 1, c = 3)$

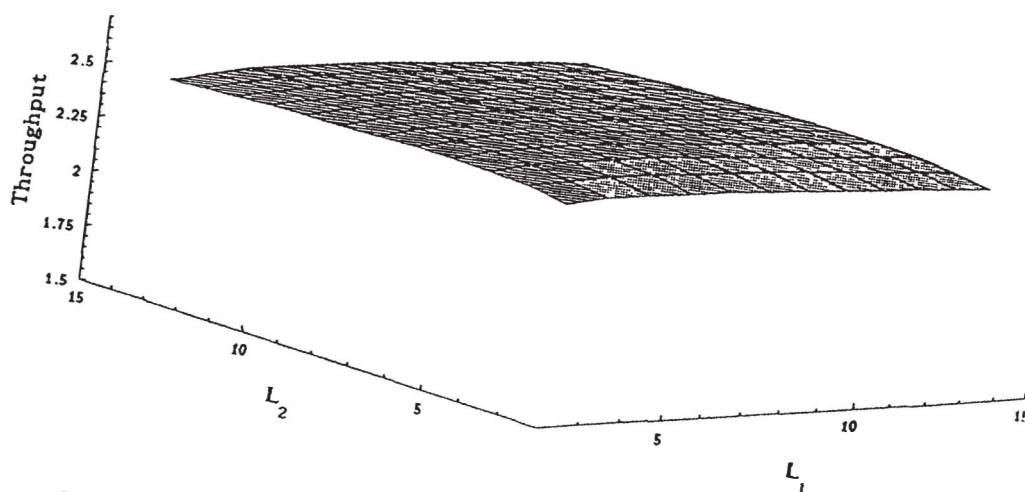


Figure 87 : Throughput versus Control
 $(C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 1.0, c_1 = 1.2, c_2 = 1.4, \mu = 1, c = 3)$

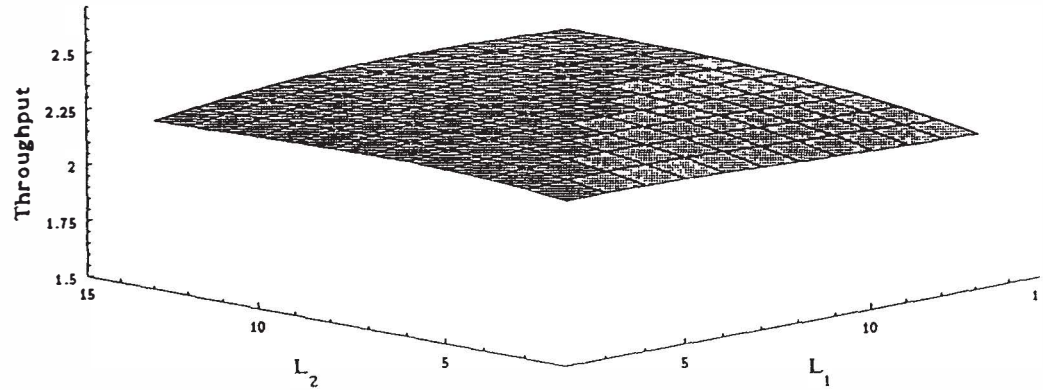


Figure 88 : Throughput versus Control

$$(C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 5.0, c_1 = 1.2, c_2 = 1.4, \mu = 1, c = 3)$$

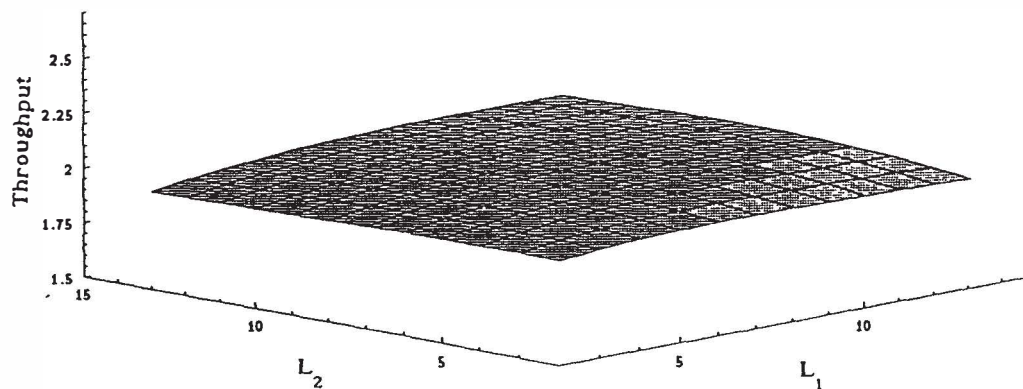


Figure 89 : Throughput versus Control

$$(C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 5.0, c_1 = 1.2, c_2 = 1.0, \mu = 1, c = 3)$$

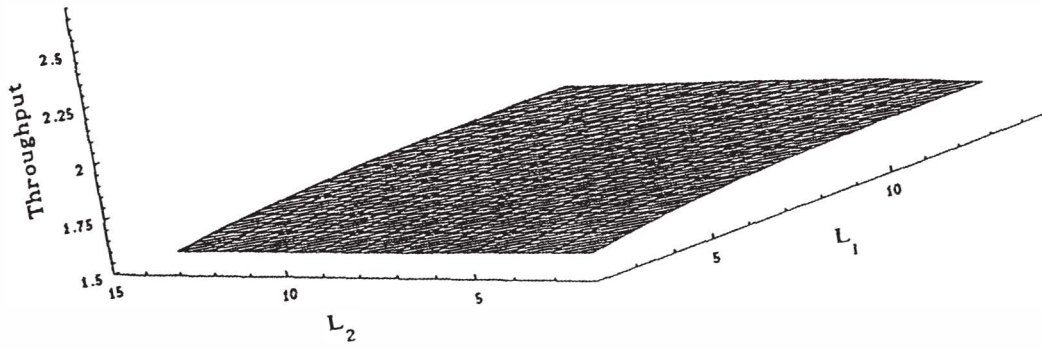


Figure 90 : Throughput versus Control
 $(C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 5.0, c_1 = 1.2, c_2 = 0.6, \mu = 1, c = 3)$

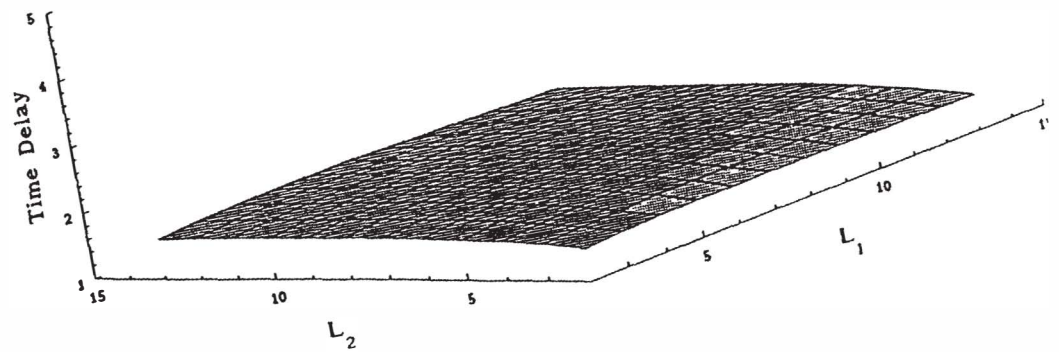


Figure 91 : Time Delay versus Control
 $(C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 0.5, c_1 = 1.2, c_2 = 1.4, \mu = 1, c = 3)$

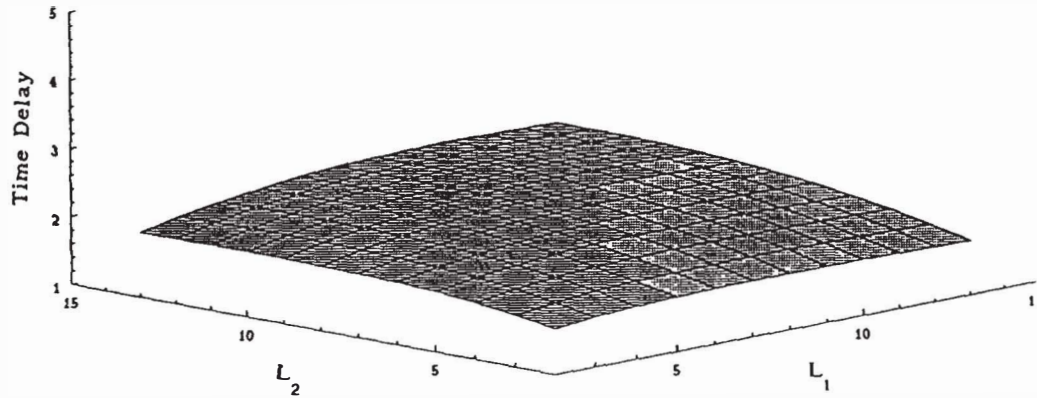


Figure 92 : Time Delay versus Control

$$(C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 1.0, c_1 = 1.2, c_2 = 1.4, \mu = 1, c = 3)$$

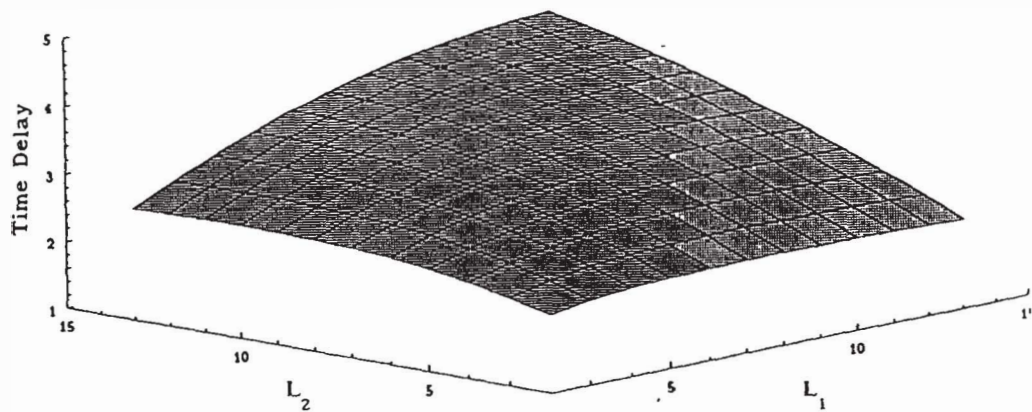


Figure 93 : Time Delay versus Control

$$(C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 5.0, c_1 = 1.2, c_2 = 1.4, \mu = 1, c = 3)$$

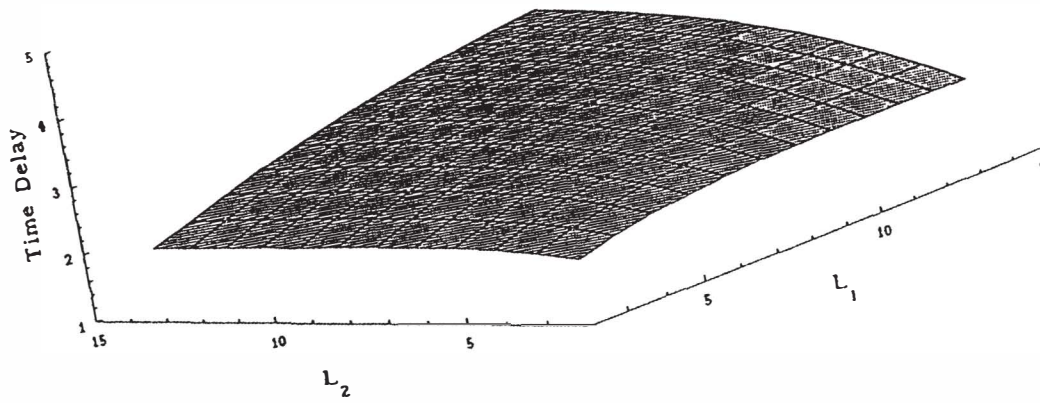


Figure 94 : Time Delay versus Control
 $(C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 5.0, c_1 = 1.2, c_2 = 1.0, \mu = 1, c = 3)$

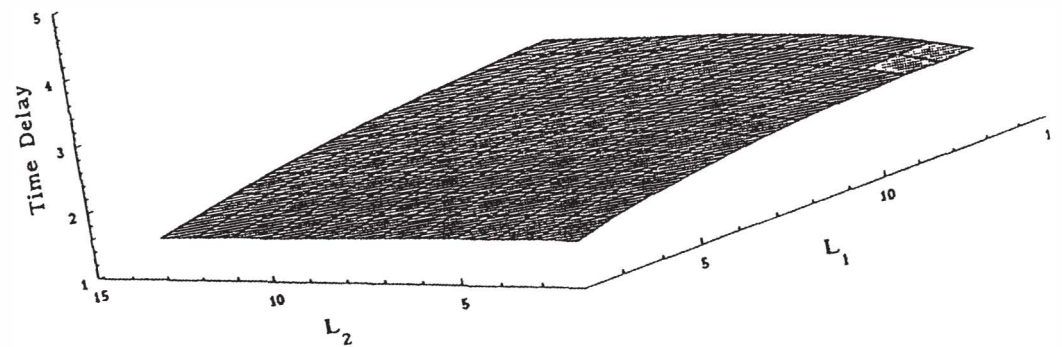


Figure 95 : Time Delay versus Control
 $(C_{\mu}^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 5.0, c_1 = 1.2, c_2 = 0.6, \mu = 1, c = 3)$

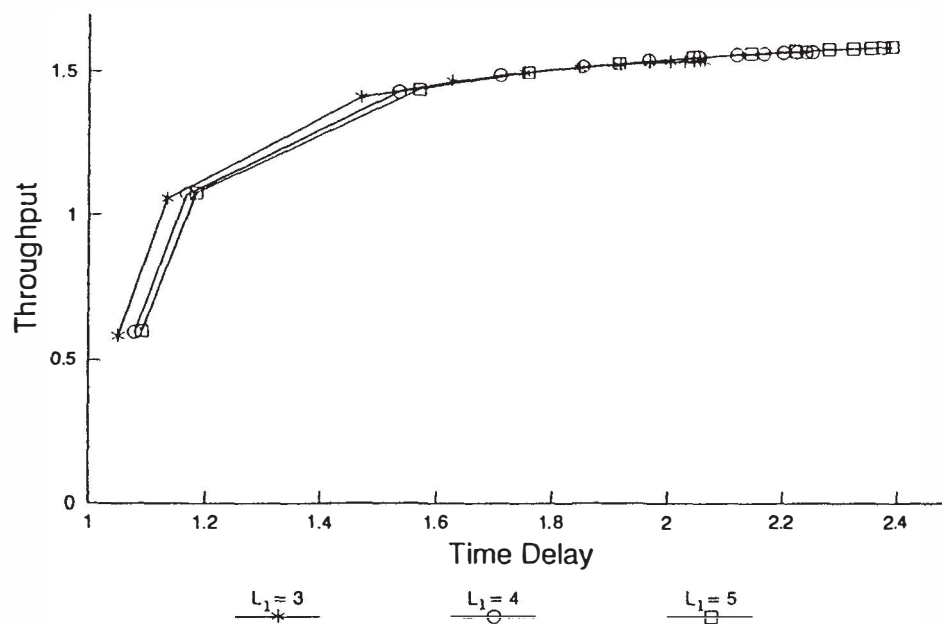


Figure 96 : Throughput versus Time Delay (L_1 is fixed,
 $C_\mu^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 1$, $\mu = 1$, $c = 3$)

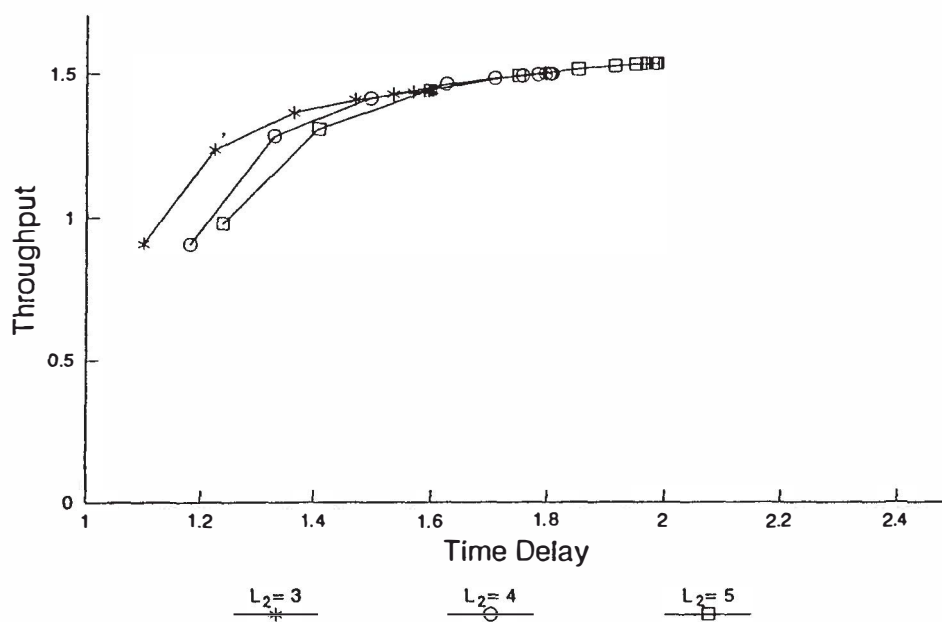


Figure 97 : Throughput versus Time Delay (L_2 is fixed,
 $C_\mu^2 = C_{\lambda_1}^2 = C_{\lambda_2}^2 = 1$, $\mu = 1$, $c = 3$)

Summary

In this chapter, the decentralised optimal flow control of a $G/G/c$ queue with K classes of users approximated by a $GE/GE/c$ queue, when the first two moments are known, has been studied. The optimal control that maximises the average throughput subject to a bounded time delay constraint has been shown to be a state dependent window type mechanism. The maximum number of packets in transit within the system (i.e. sliding window size) corresponding to a maximum throughput (when optimally controlled), can be determined from a pre-assigned upper bound on the average time delay. The direct dependence of the maximum throughput on the average time delay is also examined. Numerical examples provide useful information on how critically the optimal throughput is affected by the distributional form of the interarrival and service times. The analysis of the optimal flow control with global objectives does not favour any individual user, and depends on the relative order of their packet generation rates.

CHAPTER VIII

CONCLUSIONS AND FUTURE RESEARCH

General Summary

The motivation for this research work was the need to consider the second moments of the delay time of the controllers and the transmission time of the channels in performance analysis of heterogeneous computer communication networks. In this multiclass user environment, these important parameters should be included in stochastic analysis of the performance since various users have diverse traffic characteristics and performance requirements in utilising network resources. General closed queueing networks with multiclass users, single server and/or multiserver centres have been used in modelling the computer communication networks. In this thesis, the single server and/or multiserver general closed queueing network models with multiclass users have been analysed by using the maximum entropy approach. The approximate analytical solutions obtained are then used in the derivation of the decentralised optimal flow control of the computer communication networks with multiple controllers, single and/or multiple transmission links. The generalised exponential (GE) distributional model is chosen to approximate general distributions with known first two moments. Note that the exponential distribution is a special case of the GE distributional model.

The study can logically be divided into four parts. First, the application of general queueing networks in modelling computer system and computer communication networks followed by the problem of

controlling congestion in packet switched data communication networks was introduced (Chapter I and II). Secondly, the principle of maximum entropy and the application of Norton's theorem to large general queueing networks was presented (Chapter II) followed by the presentation of a new generalised Norton's theorem in conjunction with the maximum entropy approach for model reduction (Chapter III). Thirdly, new solutions for the queue length distributions of the single server general closed queueing networks with multiclass users (Chapter IV) and multiserver general closed queueing networks with multiclass users (Chapter V) was presented. Finally, the decentralised optimal flow control of single server general queueing networks with multiclass users is discussed in Chapter VI, and in Chapter VII the derivation of the decentralised optimal flow control of multiserver general closed queueing networks with multiclass users is presented.

The main conclusions of this research may be summarised as follows:

1. In Chapter II and III, a new unified method for model reduction has been implemented. The application of Norton's theorem for large general queueing network in conjunction with the maximum entropy approach is presented in order to simplify the complexity of the network and subsequently facilitate the analytic treatment of the derivation of optimal flow control problems.
2. The principle of maximum entropy has been used to derive new solutions for the queue length distributions of the single server general closed queueing network models with multiclass users (Chapter VI). The principle of maximum entropy has also been used to derive new approximate analytic solutions for the joint queue

length distributions of the multiserver general closed queueing networks with multiclass users (Chapter V). These solutions are implemented by making use of the GE distributional model and its pseudo memoryless properties. As a consequence, all analytical results derived can be interpreted in terms of bulk arrival and bulk service queues with geometrically distributed bulk sizes which can be represented in a one-step recursions solutions.

3. It is shown that flow control maximising the throughput of a GE type decentralised data link computer communication networks under a bounded time delay is a state dependent window type mechanism (bang-bang control) that has been traditionally used in practice. The maximum number of packets in transit within the system, L , (i.e. sliding window size), corresponding to a maximum throughput (i.e. when optimally controlled), can be determined from a preassigned upper bound on the mean time delay, the average allowed load and the parameters of underlying systems. The direct dependence of the maximum throughput on the mean time delay is also examined. The optimal flow control with global objectives results in a team decision that does not favour any individual user, and depends only on the relative order of their packet generation rates. From our studies (i.e. as shown under the Numerical Examples section), it is found that the optimal throughput and time delay are critically affected by (i) the distributional form of the interarrival and service times, and (ii) the maximum offered load.

The results in this thesis suggest that in establishing a general computer communication network, it is important (i) to take into account the variability, (ii) to have a probabilistic based facilities

and (iii) to consider the effect of resources being shared. The results described in this thesis provide a flexible tool in designing, operating or upgrading fault tolerant computer communication networks which are currently in great demand.

Suggestions for Future Research

There is an abundance of literature on issues related to queueing systems being applied to model computer communication networks and the determining of flow control in computer communication networks. Most of these issues have been analysed based on the exponential assumptions, the relaxation of which results in increasing system complexity and the need for further research.

In the analysis of the general closed queueing models, we have made the assumption that the service discipline at all the servers is FCFS except in the case of systems involving multiservers where an ample (i.e. infinite) service discipline is also included. This assumption should be relaxed by considering the different service disciplines among different classes of users which could be used to picture the priority among different source destination pairs of users in the computer communication networks.

The analysis in Chapter VI and Chapter VII can be extended further by considering the individual objectives under which each user tries to maximise its throughput subject to an upper bound on its average time delay. Hsiao and Lazar (1985) has made an attempt to study the optimal flow control of a single server exponential queueing network on individual objectives for the case of exponential type

system in which the Nash equilibrium is investigated and demonstrated to be of a particularly simple structure.

Another possible area of extension is to determine the decentralised optimal end-to-end flow control of single server and multiserver general closed queueing networks with multiclass users. Hsiao and Lazar (1990; 1991) has studied the optimal flow control for the case of exponential type system. It is shown that the network can be analysed in terms of an equivalent network by using the generalised Norton's theorem. The structure of the optimal flow control is shown to be of a window type.

The flow control scheme that has been presented is designed under the assumption that feedback information such as acknowledgement packets, is received at the controllers with delays that are negligible compared to the transmission and queueing times that the traffic encounters in the forward direction (i.e from source to destination). For the transmission and queueing delays of feedback packets, this assumption can always be made, under certain conditions (short control packets with higher priorities). However, in the last few years, it has become common knowledge that, in the emerging high-speed networks spanning over wide areas, this assumption is not valid for the constant propagation delays. New attempts will have to be made to include this delay in the flow control analysis.

Finally in conclusion, the principle of maximum entropy has proved itself as a satisfactory tool for analysing general closed queueing network models. It does not contradict the existing results of the Markovian queueing theory but extends and refines them. More research are needed to promote, strengthen and apply this technique to

other vital areas of applications such as computer integrated manufacturing systems, transportation and reliability related problems.

BIBLIOGRAPHY

- Allen, A.O. (1978) *Probability, Statistics and Queueing Theory*. New York: Academic Press.
- Almond, J. (1984) *Analytic Queueing Network Models*. M.Sc. Dissertation, University of Bradford.
- Bambos, N. and J. Walrand (1991) "On Stability and Performance of Parallel Processing Systems." *J. of the Association for Computing Machinery*. 38(2): 429-452.
- Baskett, F., K.M. Chandy, R.R. Muntz and F.G. Palacios (1975) "Open, Closed and Mixed Networks of Queues with Different Classes of Customers." *J. of the Association for Computing Machinery*. 22(2): 248-260.
- Bharath-Kumar, K. (1980) "Optimum End-to-End Flow Control in Networks." *In Conference Record of the International Conference on Communications*. pp. 23.3.1-23.3.6.
- Buzen, J.P. (1973) "Computational Algorithms for Closed Queueing Networks with Exponential Servers." *Communication of the Association of Computing Machinery*. 16(9): 529-531.
- Buzen, J.P. and A.B. Bondi (1983) "The Response Times of Priority Classes under Preemptive Resume in M/M/m Queues." *Operations Research*. 31(3): 456-465.
- Buzen, J.P. and P.J. Denning (1980) "Measuring and Calculating Queue Length Distributions." *IEEE Computer*. 13(4): 33-44.
- Chandy, K.M., U. Herzog and L. Woo (1975a) "Parametric Analysis of Queueing Networks." *IBM J. Research and Development*. 19(1): 36-42.
- Chandy, K.M., U. Herzog and L. Woo (1975b) "Approximate Analysis of General Queueing Networks." *IBM J. Research and Development*. 19(1): 43-49.
- Courtois, P.J. (1977) *Decomposability: Queueing and Computer System Applications*. New York: Academic Press.
- Davies, D.W. (1972) "The Control of Congestion in Packet-Switching Networks." *IEEE Transactions on Communications*. 20(3): 546-550.
- De Smit, J.H.A. (1973a) "On the Many Server Queue with Exponential Service Times." *Advanced Applied Probability*. 5: 170-182.
- De Smit, J.H.A. (1973b) "Some General Results for Many Server Queues." *Advanced Applied Probability*. 5: 153-169.

- El-Affendi, M.A. and D.D. Kouvatsos (1983) "A Maximum Entropy Analysis of the M/G/1 and G/M/1 Queueing Systems at Equilibrium." *Acta Informatica*. 19: 339-355.
- Foxley, E. and O. Salman (1981) "Validation of an Analytic Model of Computer Performance." *The Computer J.* 24(4): 347-352.
- Gelenbe, E. and G. Pujolle (1976) "The Behaviour of a Single Queue in a General Queueing Network." *Acta Informatica*. 7: 123-136.
- Gelenbe, E. and Z. Liu (1988) "Performance Analysis Approximations for Parallel Processing in Multiprocessor Systems." In Proceedings edited by M. Cosnard, M.H. Barton and M. Vanneschi: Parallel Processing, pp. 363-375.
- Gordon, J.J. (1990) "The Evaluation of Normalizing Constants in Closed Queueing Networks." *Operations Research*. 38(5): 863-869.
- Guillon, A. (1977) "A Case Study of Performance Modeling a Large Computer System." In Proceedings edited by K.M. Chandy and M. Reiser: Computer Performance, pp. 235-248.
- Harrison, P.G. (1985) "On Normalizing Constants in Queueing Networks." *Operations Research*. 33(3): 464-468.
- Hokstad, P. (1986) "Relations for the Workload of the GI/G/s Queue." *Advanced Applied Probability*. 17: 887-904.
- Hsiao, M.T. and A.A. Lazar (1984) "Bottleneck Modelling and Decentralised Optimal Flow Control - I. Global Objectives." In Proceedings of the Eighteenth Conference on Information Sciences and Systems, New Jersey, 13-15 March, 1984, pp. 169-173.
- Hsiao, M.T. and A.A. Lazar (1985) "Bottleneck Modelling and Decentralised Optimal Flow Control - II. Individual Objectives." In Proceedings of the Nineteenth Conference on Information Sciences and Systems, Maryland, 27-29 March, 1985, pp. 558-563.
- Hsiao, M.T. and A.A. Lazar (1990) "Optimal Flow Control of Multiclass Queueing Networks with Partial Information." *IEEE Transaction on Automatic Control*. 35(7): 855-860.
- Hsiao, M.T. and A.A. Lazar (1991) "Optimal Decentralised Flow Control of Markovian Queueing Networks with Multiple Controllers." *Performance Evaluation*. 13(3): 181-204.
- Iglehart, D.L. and G.S. Shedler (1978) "Simulation of Response Times in Finite-Capacity Open Networks of Queues." *Operations Research*. 26(5): 896-914.
- Jaynes, E.T. (1957) "Information Theory and Statistical Mechanics." *Physical Review*. 106(4): 620-630.

- Kelton, W.D. and A.M. Law (1985) "The Transient Behaviour of the M/M/s Queue, with Implication for Steady-State Simulation." *Operations Research*. 33(2): 378-396.
- Kumira, T. (1991a) "Refining Cosmetatos, Approximation for the Mean Waiting Time in the M/D/s Queue." *J. Operational Research Society*. 42(7): 595-603.
- Kimura, T. (1991b) "Approximating the Mean Waiting Time in the GI/G/s Queue." *J. Operational Research Society*. 42(11): 959-970.
- Kimura, T. (1991c) "Approximations for the Waiting Time in the GI/G/s Queue." *J. of the Operations Research Society of Japan*. 34(2): 173-186.
- Kimura, T. (1992) "Interpolation Approximations for the Mean Waiting Time in a Multi-Server Queue." *J. of the Operations Research Society of Japan*. 35(1): 77-92.
- Klienrock, L. (1975a) *Queueing Systems Volume I: Theory*. New York: John Wiley Inc.
- Klienrock, L. (1975b) *Queueing Systems Volume II: Computer Applications*. New York: John Wiley Inc.
- Klienrock, L. (1979) "Power and Deterministic Rules of Thumb for Probabilistic Problems in Computer Communications." *In Conference Record of the International Conference on Communications*. pp. 43.1.1-43.1.10.
- Knessl, C., B.J. Matkowsky, Z. Schuss and C. Tier (1990) "An Integral Equation Approach to the M/G/2 Queue." *Operations Research*. 38(3): 506-518.
- Kobayashi, H. (1974) "Application of the Diffusion Approximation to queueing Networks I: Equilibrium Queue Distributions." *J. of the Association for Computing Machinery*. 21(2): 316-328.
- Kouvatsos, D. D. (1984) "Maximum Entropy: A Bridge between Stochastic and Operational Analysis." Research Report. CCr71, University of Bradford.
- Kouvatsos, D.D. (1985) "Maximum Entropy Methods for General Queueing Networks." *In Proceedings edited by D. Potier: Modelling Techniques and Tools for Performance Analysis*, pp. 589-608.
- Kouvatsos, D.D. (1986a) "A Maximum Entropy Queue Length Distribution for a G/G/1 Finite Capacity Queue." *Performance Evaluation Review*. 4: 224-236.
- Kouvatsos, D.D. (1986b) "A Universal Maximum Entropy Algorithm for the Analysis of General Closed Networks." *In Proceedings edited by T. Hasegawa, H Takagi and Y. Takahashi: Computer Networking and Performance Evaluation*, pp. 113-124.

- Kouvatsos, D.D. (1988a) "A Maximum Entropy Analysis of the G/G/1 Queue at Equilibrium." *J. Operational Research Society*. 39(2): 183-200.
- Kouvatsos, D.D. (1988b) "Maximum Entropy Two-Station Cyclic Queues with Multiple General Servers." *Acta Informatica*. 26: 241-267.
- Kouvatsos, D.D. (1989b) "A Universal Maximum Entropy Algorithm for General Multiple Class Open Networks with Mixed Service Disciplines." In Proceedings edited by R. Puigjaner and D. Potier: Modelling Techniques and Tools for Computer Performance Evaluation, pp. 397-419.
- Kouvatsos, D.D. and A.T. Othman (1989) "Optimal Flow Control of a G/G/1 Queue." *International J. of Systems Science*. 20(2): 251-265.
- Kouvatsos, D.D. and A.T. Othman (1989b) "Optimal Flow Control of a G/G/c Finite Capacity Queue." *J. Operational Research Society*. 40(7): 659-670.
- Kouvatsos, D.D. and A.T. Othman (1989c) "Optimal Flow Control of an End-to-end Communication Network with Fixed Routing." *International J. of Systems Science*. 20(8): 1419-1430.
- Kouvatsos, D.D. and A.T. Othman (1989d) "Optimal Flow Control of an End-to-end Communication Network with Random Routing." *IEE Proceedings*. Part E-136(2): 90-100.
- Kouvatsos, D.D. and J. Almond (1988b) "Maximum Entropy Two-Station Cyclic Queues with Multiple General Servers." *Acta Informatica*. 26: 241-267.
- Kouvatsos, D.D. and N.M. Tabet-Aouel (1989) "A Maximum Entropy Priority Approximation for a Stable G/G/1 Queue." *Acta Informatica*. 27: 247-286.
- Kouvatsos, D.D. and N.M. Tabet-Aouel (1990) "Product-Form Approximations for an Extended Class of General Closed Queueing Networks." In Proceedings of the 14th IFIP WG 7.3 International Symposium on Computer Performance Modelling, Measurement and Evaluation, 12-14 September 1990. Edinburgh, Scotland, pp. 301-315.
- Ku-Mahamud, K.R. and A.T. Othman (1990) "Norton-Maximum Entropy Methods for General Queueing Networks." *J. of the Singapore National Academy of Science*. 18: 140-149.
- Ku-Mahamud, K.R. and A.T. Othman (1990) "Model Reduction of General Queueing Networks." *International J. of Systems Science*. 24(1): 183-192.

- Ku-Mahamud, K.R and A.T. Othman (1992) "A One-Step Maximum Entropy Recursions of Decentralised Control General Data Communication Networks." *In Proceedings of the Franco-Seams Conference on The Application of Modern Mathematics in Engineering Related Problems*, Bandung, Indonesia, 22-24 June, 1992.
- Lam, S.S. (1978) "A New Measure for Characterizing data Traffic." *IEEE Transactions on Communications*. 26(1): 137-140.
- Lam, S.S. (1979) "Congestion Control of Store-and-Forward Networks by Input Buffer Limits - An Analysis." *IEEE Transactions on Communications*. 27(1): 127-133.
- Lam, S.S. and Y.L. Lien (1981) "Modelling and Analysis of Flow Controlled Packet Switching Networks." *In Proceedings of Data Communication Symposium*, Mexico City, Mexico.
- Lazar, A.A. (1982a) "Centralised Optimal Control of a Jacksonian Network." *In Proceedings of the Sixteenth Conference on Information Science and Systems*, 17-19 March 1982, New Jersey, pp. 316.
- Lazar, A.A. (1982b) "Data Link Flow Control with Partial Observations." *In Twentieth Annual Allerton Conference on Communication, Control and Computing*, 6-8 October, 1982, Urbana-Champaign.
- Lazar, A.A. (1983a) "The Throughput Time Delay Function of an M/M/1 Queue." *IEEE Transactions on Information Theory*. 29(6): 914-918.
- Lazar, A.A. (1983b) "Optimal Flow Control of a Class of Queueing Networks in Equilibrium." *IEEE Transactions on Automatic Control*. 28(11): 1001-1007.
- Lazar, A.A. (1984) "Optimal Flow Control of an M/M/m Queue." *J. of the Association for Computing Machinery*. 31(1): 86-98.
- Marsan, M.A. and F. Gregoretti (1981) "Memory Interference Models for a Multi-Microprocessor System with a Shared Bus and a Single External Common Memory." *Euromicro J.* pp. 124-133.
- Othman, A.T., K.R. Ku-Mahamud and M.H. Selamat "Analytical Analysis of Decentralised Control General Data Communication Networks." *In Proceedings of the 10th Symposium on Computational Statistics*, Neuchatel, Switzerland, August 1992, pp. 317-322.
- Pawlikowski, K. (1990) "Steady-State Simulation of Queueing Processes: A Survey of Problems and Solutions." *Association of Computing Machinery Computing Surveys*. 22(2): 123-170.
- Pawlita, P.F. (1981) "Traffic Measurements in Data Networks, Recent Measurement Results, and Some Implications." *IEEE Transactions on Communications*. 29(4): 525-535.

- Pennotti, M.C. and M. Schwartz (1975) "Congestion Control in Store and Forward Tandem Links." *IEEE Transactions on Communications*. 23(12): 1434-1443.
- Pujolle, G. and C. Soula (1979) "A Study of Flows in Queueing Networks and an Approximate Method for Solution." *In Proceedings* edited by M. Arato, A. Butrimenko and E. Gelenbe: Performance of Computer Systems, pp. 375-389.
- Reiser, M. and H. Kobayashi (1974a) "Accuracy of the Diffusion Approximation for Some Queueing Systems." *IBM J. Research and Development*. 18(2): 110-124.
- Reiser, M. and H. Kobayashi (1974b) "Queueing Networks with Multiple Closed Chains: Theory and Computational Algorithms." *IBM J. Research and Development*. pp. 283-294.
- Reiser, M. and S.S. Lavenberg (1980) "Mean-Value Analysis of Closed Multichain Queueing Networks." *J. of the Association for Computing Machinery*. 27(2): 313-322.
- Robertazzi, T.G. and A.A. Lazar (1985) "On the Modelling and Optimal Flow Control of the Jacksonian Network." *Performance Evaluation*. 5:29-43.
- Russell, E.C. (1989) *Building simulation Models with SIMSCRIPT II.5*. California: CACI Products Company.
- Sauer, C.H. and K.M. Chandy (1975) "Approximate Analysis of Central Server Models." *IBM J. Research and Development*. 19: 301-313.
- Sauer, C.H. and K.M. Chandy (1981) *Computer Systems Performance Modelling*. New Jersey: Prentice Hall.
- Sevcik, K.C., A.I. Levy, S.K. Tripathi and J.L. Zahorjan (1977) "Improving Approximations of Aggregate Queueing Networks Subsystems." *In Proceedings* edited by K.M. Chandy and M. Reiser: Computer Performance. pp. 1-22.
- Schwartz, M. (1980) "Routing and Flow Control in Data Networks." Res. Rep. RC 8353 (#36319), IBM Thomas J. Watson Research Centre, Yorktown Heights, New York, pp. 273-301.
- Schwartz, M. (1987) *Telecommunication Networks: Protocols, Modeling and Analysis*. New York: Addison Wesley.
- Shannon, C.E. (1948) "A Mathematical Theory of Communication." *Bell System Tech. J.* 27: 379-423.
- Shore, H. (1988a) "Simple Approximations for the GI/G/c Queue-I: The Steady-State Probabilities." *J. Operational Research Society*. 39(3): 279-284.

- Shore, H. (1988b) "Simple Approximations for the GI/G/c Queue-II: The Moments, the Inverse Distribution Function and the Loss Function of the Number in the System and of the Queue Delay." *J. Operational Research Society*. 39(4): 381-391.
- Shore, J.E. and R.W. Johnson (1980) "Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy." *IEEE Transactions on Information Theory*. IT-26: 26-37.
- Smith, C.E. (1949) *Communication Circuit Fundamentals*. New York: McGraw Hill.
- Tabet-Aouel, N. and D.D. Kouvatsos (1992) "On an Approximation to the Mean Response Times of Priority Classes in a Stable G/G/c/PR Queue." *J. Operational Research Society*. 43(3): 227-239.
- Takacs, L (1962) *Introduction to the Theory of Queues*. New York: Oxford University Press.
- Woodside, C.M., D.A. Stanford and B. Pagurek (1984) "Optimal Prediction of Queue Lengths and Delays in GI/M/m Multiserver Queues." *Operations Research*. 32(4): 809-817.
- Wu, J. -S. and W.C. Chan (1989) "Maximum Entropy Analysis of Multiple Server Queueing Systems." *J. of the Operational Research Society*. 40(9): 815-825.
- Yao, D.D. (1985) "Refining the Diffusion Approximation for the M/G/m Queue." *Operations Research*. 33(6): 1266-1277.

APPENDICES

APPENDIX A

The Generalised Exponential (GE) Distribution Model

The GE probability density function (pdf) of the form (cf. 1.1)

$$\phi(t) = \left(\frac{C^2 - 1}{C^2 + 1} \right) u_0(t) + \frac{4v}{(C^2 + 1)^2} \exp\left(-\frac{2vt}{C^2 + 1}\right), \quad t \geq 0, \quad [A1]$$

where $1/v$ is the mean value, C is the coefficient of variation and $u_0(t)$ is the unit impulse function, is an 'extremal' member of a family of two-phase models with the same first two moments and a pdf of the universal form

$$f(t, k) = \alpha_1(k) v_1(k) \exp\{-v_1(k)t\} + \alpha_2(k) v_2(k) \exp\{-v_2(k)t\}, \quad t \geq 0, \quad [A2]$$

such that

$$\frac{\alpha_1(k)}{v_1(k)} + \frac{\alpha_2(k)}{v_2(k)} = \frac{1}{v}, \quad \alpha_1(k) + \alpha_2(k) = 1, \quad [A3]$$

$$2v^2 \left(\frac{\alpha_1(k)}{v_1(k)^2} + \frac{\alpha_2(k)}{v_2(k)^2} \right) - 1 = C^2 \quad [A4]$$

and

$$\lim_{k \rightarrow +\infty} f(t, k) \rightarrow \phi(t), \quad t \geq 0,$$

$$k \rightarrow +\infty (C^2 > 1)$$

or

$$k \rightarrow -\infty (C^2 < 1) \quad [A5]$$

where $1/v_1(k)$, $1/v_2(k)$ are the mean values of the two service phases respectively, and $\{k\}$ is a set of 'tuning' parameters. $\{k\}$ can be used to generate a family of two-phase hyperexponential (H_2) distributions with parameters $\{v, C^2 > 1\}$, with phase-selection probabilities $\{\alpha_1(k) > 0, \alpha_2(k) > 0\}$ and $k \in (1, +\infty)$, or 'special', generally 'improper' (or pseudo) hypoexponential (h_2) models with parameters $\{v, C^2 < 1\}$, with phase-selection indexes $\{\alpha_1(k) < 0, \alpha_2(k) > 1\}$ and $k \in (-\infty, 0)$ satisfying (in both cases) the relations

$$v_1(k) = k\alpha_1(k)v, \quad [A6]$$

$$v_2(k) = \frac{k\alpha_2(k)v}{k-1}, \quad [A7]$$

Note that for $C^2 > 1$, the GE pdf [A1] has been shown (Kouvatsos 1988a) to maximise the entropy function

$$H\{f(t)\} = - \int_0^x f(t) \log \{f(t)\} dt, \quad [A8]$$

subject to the mean value constraints (i.e. which based on the first two moments):

$$\int_0^x f(t) dt = 1, \quad [A9]$$

$$\int_0^x \xi(t) f(t) dt = \frac{C^2 - 1}{C^2 + 1}, \quad \xi(t) = \begin{cases} 0 & \text{if } t \leq 0, \\ 1 & \text{if } t > 0, \end{cases} \quad [\text{A10}]$$

$$\int_0^x t f(t) dt = \frac{1}{\mu}, \quad [\text{A11}]$$

respectively. In information theoretic terms (Jaynes 1968) the GE probability density function can be interpreted as the minimally biased distribution subject to given information expressed by the constraints [A9] - [A11]. In this case the GE model [A1] can be interpreted as a limiting case of an H_2 distribution where one of the two phases has zero service time. Moreover, the underlying counting (e.g. arrival process of a GE inter-event time is equivalent to a compound Poisson process (M^B) with parameter $= 2\nu/(1+C^2)$ and geometrically distributed batches with mean $= (1+C^2)/2$ and squared coefficient of variation $= (C^2-1)/(C^2+1)$. For $C^2 < 1$, the 'special' h_2 model of the form [A2] may produce, in certain cases, an 'improper' distribution when the probability distribution function

$$F(x, k) = \int_0^x f(t, k) dt, \quad x \geq 0, \quad [\text{A12}]$$

produces negative values for a small range of x values close to zero, especial for $C^2 < 0.5$ and $k < -1$. For $C^2 = 0.5$, the h_2 model is identical to a two-phase Erlang (E_2) distribution at $k \rightarrow 0^-$ (0^- indicates a negative value close to zero, e.g. -0.001) and is very close to a 'proper' classical hypoexponential-2 distribution for $0.5 < C^2 < 1$ at $k \rightarrow 0^-$. Although the GE model of [A1] is improper for $C^2 < 1$, since

$$\lim_{k \rightarrow -\infty} \Phi(0,k) = \frac{C^2 - 1}{C^2 + 1} < 0, \quad [A13]$$

nevertheless it represents, in general, a very useful and robust overall approximation of general distributions (El-Affendi and Kouvatsos 1983, Kouvatsos 1985 , 1986a, 1986b). The utility of other improper two-phase distributions with $C^2 > 1$ in systems performance modelling has also been pointed out in Sauer (1975). However, note that closed-form GE-type solutions of isolated queues and networks are simpler to derive owing to the psuedo memoryless properties of the GE distributional model.

APPENDIX B

A Universal Maximum Entropy Algorithm

Consider an arbitrary single class general closed networks as described in Chapter III. The maximum entropy algorithm is implemented as follows (Kouvatsos 1986b):

Input parameters

- M - number of single server centres
- N - number of customers in the closed network
- μ_i - service rates at centre i
- C_{si} - coefficient of variation of service times at centre i
- q_{ij} - transition probabilities

Step 1 {Part 1: The iteration routine for the 'pseudo' open network.}

Step 1.1 {Obtain the invariants in the iteration.}

Step 1.1.1 {Solve the job flow balance equations [3.4] by using the Gauss-Seidal equation to obtain λ_i , $i = 1, 2, \dots, M$.}

Step 1.1.2 {Obtain aggregate parameters.}

for i := 1 to M do {step 1.1.2}

Step 1.1.2.1

$$\rho_i \leftarrow \lambda_i / \mu_i$$

Step 1.1.2.2

for j := 1 to M do

$$\lambda_{ij} \leftarrow \lambda_i q_{ij}$$

Step 1.1.3 {Obtain network bottleneck b: $\rho_b \leftarrow \max_i \{\rho_i\}$ }

Step 1.2 {Initialise flow coefficient of variation}

for $i := 1$ to M do $\{C_{di}^2 \leftarrow C_{si}^2, C_{ai}^2 \leftarrow 1\}$

Step 1.2.1 {Initialise multiplicative constant:

$$\phi \leftarrow (1/\rho_b) - (0.01/\rho_b)\}$$

for $i := 1$ to M do $\{\phi_i \leftarrow \phi \rho_i\}$

Step 1.3 {Iteration: repeat until convergence of C_{di}^2 and C_{ai}^2 }

Step 1.3.1 {Apply the Newton-Raphson iteration method}

Step 1.3.1.1

for $i := 1$ to M do

$$\langle n_i \rangle \leftarrow \frac{\phi \rho_i}{2} \left\{ 1 + \frac{C_{ai}^2 + \phi \rho_i C_{si}^2}{1 - \phi \rho_i} \right\}$$

Step 1.3.1.2 {Solve w.r.t ϕ the equation-constraint to

obtain $\phi' \leftarrow \phi$ }

$$\sum_{i=1}^M \langle n_i \rangle - N = 0$$

Step 1.3.1.3

for $i := 1$ to M do

$\phi_i \leftarrow \phi' \rho_i$, check $\phi_i < 1$

Step 1.3.2 {Calculate new flow squared coefficients of variation}

Step 1.3.2.1 {Inter-departure process}

for $i := 1$ to M do

$$C_{di}^2 \leftarrow \phi_i (1 - \phi_i) + (1 - \phi_i) C_{ai}^2 + \phi_i^2 C_{si}^2$$

Step 1.3.2.2 {Split process}

for $i := 1$ to M do

$$C_{dji}^2 \leftarrow 1 + q_{ji} (C_{dj}^2 - 1)$$

Step 1.3.2.3 {Merging process}

for $i := 1$ to M do

$$C_{ai}^2 \leftarrow \frac{\sum_{j=1}^K \lambda_i q_{ji}}{\sum_{j=1}^K \lambda_i q_{ji} / \{C_{dji}^2 + 1\}} - 1$$

Step 1.4 {Compute Lagrangian coefficients}

$$\begin{aligned} & \underline{\text{for } i := 1 \text{ to } M \text{ do}} \\ & \quad g_i \leftarrow \frac{\phi_i^2}{(1 - \phi_i)(\langle n_i \rangle - \phi_i)} \\ & \quad x_i \leftarrow \frac{\langle n_i \rangle - \phi_i}{\langle n_i \rangle} \\ & \quad g_{0i} \leftarrow 1 \end{aligned}$$

Step 2 {Part 2: The convolution routine for the (original) closed network}

Step 2.1 {Normalising constant $Z(p)$ }

Step 2.1.1

$$\begin{aligned} & \underline{\text{for } i := 1 \text{ to } M \text{ do}} \\ & \quad g_i \leftarrow g_{0i} g_i \end{aligned}$$

Step 2.1.2 {Initialise Z }

$$Z(k) \leftarrow \begin{cases} 1 & \text{for } k = 0, \\ x_1 g_1, & \text{for } k = 1, \\ x_1 Z(k-1), & \text{for } k = 2, 3, \dots, N, \end{cases}$$

Step 2.1.3 {Convolute for Z } { $z(0) \leftarrow 0$ }

$$\begin{aligned} & \underline{\text{for } i := 1 \text{ to } M \text{ do}} \\ & \quad \underline{\text{for } k := 1 \text{ to } N \text{ do}} \\ & \quad \quad z(k) \leftarrow Z(k) \\ & \quad \quad Z(k) \leftarrow z(k) + x_i [(g_i - 1) z(k-1) + Z(k-1)] \end{aligned}$$

Step 2.2 {Performance statistics}

Function (M)

for $i := 1$ to $M-1$ do
for $k := 1$ to K do
 $z(k) \leftarrow Z(k) - x_i[(g_i - 1) z(k-1) + Z(k-1)]$

Function (i)

where Function (i) is:

$$f_i(k) \leftarrow \begin{cases} 1 & \text{for } k = 0, \\ x_i g_i, & \text{for } k = 1, \\ x_i f_i(k-1), & \text{for } k = 2, 3, \dots, N, \end{cases}$$

$$\langle n_i \rangle_K \leftarrow \sum_{k=1}^K k f_i(k) z(K-k)/Z(K)$$

{mean queue length for centre i}

$$U_i \leftarrow 1 - z(K)/Z(K) \text{ {utilisation of server i}}$$

Step 2.2.1 {Job flow balance correction}

for $i := 1$ to M do
if $\rho_i/U_i \neq \text{constant}$ then
 $g_{0i} \leftarrow \rho_i g_{0i} K \left\{ U_i \sum_{j=1}^M \langle n_j \rangle_K \rho_j / U_j \right\}^{-1}$
 Return to step 2 until $\rho_i/U_i = \text{constant } \forall i$

Step 2.2.2

for $i := 1$ to M do
 $X_i \leftarrow U_i \mu_i$ (mean throughput of centre i)

Step 2.3 {Feedback correction}

for $i := 1$ to M do
if $q_{ii} > 0$ then
 $\lambda_i \leftarrow (1 - q_{ij}) \lambda_i$
 $\lambda_{ii} \leftarrow 0$
 $C_{si}^2 \leftarrow + (1 - q_{ii}) C_{si}^2$
for $j := 1$ to M do
if $i \neq j$ then $q_{ij} \leftarrow q_{ij} / (1 - q_{ii})$

$$q_{11} \leftarrow 0$$

2.3.1 {Go to step 1.3}

Otherwise STOP

This algorithm has efficiently been implemented in C programming language. Earlier versions of the algorithm are available in Pascal (Almond 1984) and FORTRAN 77 (Kouvatsos 1984).

VITA

The author was born on the 15th November, 1960 in Jitra, in the state of Kedah, Malaysia. She completed her primary education from Sekolah Rendah Sultan Ahmad Tajuddin, Jitra and her secondary education from Sekolah Menengah Sains Kedah, Sungai Petani and Sekolah Menengah Sains Perlis, Kangar. She received her 'A' Level Certificate and Bachelor of Science in Mathematical Sciences from Norwich City College, England and University of Bradford, England in 1980 and 1983 respectively. She received her Postgraduate Diploma in Education from Universiti Kebangsaan Malaysia in 1984.

The author joined Sekolah Menengah Panglima Bukit Gantang, Parit Buntar, as a teacher in April, 1984 and in August, of the same year she joined Universiti Utara Malaysia as a tutor. She received her Master of Science in Computing from the University of Bradford, England in 1986 and was later appointed as a lecturer in January, 1987. At the moment, the author is attending a Ph. D. programme in Computer Science under the sponsorship of Universiti Utara Malaysia and Public Services Department.

I certify that the Board of Examiners have met on December 29, 1993, to conduct the final examination of Ku Ruhana Ku Mahamud on her Doctor of Philosophy dissertation entitled "Analysis and Decentralised Optimal Flow Control of Heterogeneous Computer Communication Network Models" in accordance with Universiti Pertanian Malaysia (Higher Degree) Act 1980 and Universiti Pertanian Malaysia (Higher Degree) Regulations 1981, and the Board recommended that the candidate be awarded the relevant degree. The Board of Examiners for the candidate consists of the following persons:

AUB TALIB OTHMAN, Ph.D.
Associate Professor/Head
Department of Computer Science
Faculty of Science and Environmental Studies
(Chairman)

ALI MAMAT, Ph.D.
Department of Computer Science
Faculty of Science and Environmental Studies
(Internal Examiner)

BORHANUDDIN MOHD. ALI, Ph.D.
Associate Professor/Head
Department of Electronics and Computer Engineering
Faculty of Engineering
(Internal Examiner)

MASHKURI HJ. YAACOB, Ph.D.
Professor/Director
Computer Centre
University of Malaya
(External Examiner)

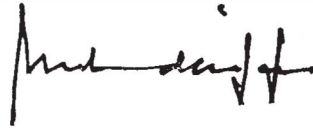
PETER G. HARRISON, Ph.D.
Department of Computing
Imperial College of Science, Technology and
Medicine
(External Examiner).



MOHD. ARIFF HUSSEIN, Ph.D.
Professor/Dean of Graduate School
Universiti Pertanian Malaysia.

Date: 17 JAN 1994

This dissertation was submitted to the Senate of Universiti Pertanian Malaysia and was accepted as fulfilment of the requirements for the degree of Doctor of Philosophy.



MOHD. ARIFF HUSSEIN, Ph.D.
Professor/Dean of Graduate School
Universiti Pertanian Malaysia.

Date: 10 MAR 1994