OPTIMIZATION OF WORKLOAD ALLOCATION PROBLEM IN A NETWORK OF HETEROGENEOUS COMPUTER SYSTEMS

A thesis submitted in full fulfilment of the requirements for the degree of Doctor of Philosophy in the Faculty of Information Technology, Universiti Utara Malaysia

By

Rahela Abdul Rahim

Pusat Pengajian Siswazah

(Centre for Graduate Studies)
Jabatan Hal Ehwal Akademik
(Department of Academic Affairs)
Universiti Utara Malaysia

PERAKUAN KERJA TESIS / DISERTASI

(Certification of thesis / dissertation)

Saya, yang bertandatangan, memperakukan bahawa
(I, the undersigned, certify that)

RAHELA ABDUL RAHIM

calon untuk Ijazah **DOKTOR FALSAFAH (Ph.D.)**
(candidate for the degree of)

telah mengemukakan tesis / disertasi yang bertajuk
(has presented his/her thesis / dissertation of the following title)

"OPTIMIZATION OF WORKLOAD ALLOCATION PROBLEIM IN A NETWORK OF HETEROGENEOUS COMPUTER SYSTEMS"

seperti yang tercatat di muka surat tajuk dan kulit tesis / disertasi
(as it appears on the title page and front cover of thesis / dissertation)

bahawa kertas projek tersebut boleh diterima dari segi bentuk serta kandungan dan meliputi bidang ilmu dengan memuaskan, sebagaimana yang ditunjukkan oleh calon dalam ujian lisan yang diadakan pada : **26 JUN 2005**
*that the project paper acceptable in the form and content and a satisfactory knowledge of the field is covered by the thesis, was demonstrated by the candidate through an oral examination held on : **26 JUNE 2005***

Pengerusi Viva *(Chairman for Viva)*	: **Prof. Madya Dr. Mohd Zaini Abd Karim**	Tandatangan *(Signature)*
Penilai Luar *(External Assessor)*	: **Prof. Dr. Mohammad Ishak Desa**	Tandatangan *(Signature)*
Penilai Dalaman *Principal Assessor)*	: **Prof. Madya Dr. Razman Mat Tahar**	Tandatangan *(Signature)*
Penyelia Utama *(Principal Supervisor)*	: **Prof. Dr. Hajah Ku Ruhana Ku Mahamud**	Tandatangan *(Signature)*

Tarikh:
(Date)

PERMISSION TO USE

In presenting this thesis in full fulfilment of the requirements for a post-graduate degree from Universiti Utara Malaysia, I agree that the University Library may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by my supervisor or, in their absence, by the Director of Graduate Studies. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Requests for permission to copy or to make other use of materials in this thesis, in whole or in part, should be addressed to:

Director of Graduate Studies
Universiti Utara Malaysia
06010 UUM Sintok
Kedah Darul Aman

ABSTRAK

PENGOPTIMUMAN MASALAH PERUNTUKAN BEBAN KERJA DALAM PERSEKITARAN SISTEM KOMPUTER-KOMPUTER HETEROGEN YANG BERANGKAIAN

Model Berbilang gilir Berbilang Pelayan dalam Teori Baris gilir digunakan untuk memodel peruntukan beban kerja yang optimum dengan satu kelas dan berbilang kelas bagi kerja kepada sekumpulan komputer yang selari diperkenalkan diikuti oleh masalah menentukan saiz penimbal yang optimum dan berkait dengan ketibaan beban kerja ke suatu komputer. Model taburan eksponen teritlak (GE) dengan mengetahui dua momen pertama digunakan untuk mewakili taburan-taburan masa antara ketibaan dan servis di mana setiap kerja mempunyai ciri-ciri trafik yang pelbagai. Model-model lain bagi taburan servis seperti eksponen, Erlang-k dan Gamma juga digunakan untuk mengembangkan keupayaan kerja yang dicadangkan. Satu algoritma baru bagi peruntukan beban kerja menggunakan disiplin masuk dulu keluar dulu bergabung dengan pengoptimuman sistem baris gilir eksponen teritlak dicadangkan untuk meminimakan masa tindakbalas dalam suatu rangkaian komputer-komputer. Algoritma ini mempunyai kelebihan berbanding skim peruntukan beban kerja klasik dan dibuat perbandingan. Ukuran-ukuran prestasi, purata kepanjangan baris gilir dan purata masa tindakbalas bagi skim yang dicadangkan secara praktikalnya menunjukkan kemajuan.

Prinsip pengoptimuman dan model taburan eksponen teritlak digunakan untuk menghasilkan suatu model peruntukan beban kerja yang baru bagi satu kelas kerja dalam rangkaian sistem baris gilir. Konsep ulangpakai dicadangkan untuk

mendapatkan penyelesaian bagi menentukan peruntukan beban kerja individu dalam persekitaran berbilang kelas. Kajian ini menarik minat di mana kedua-dua andaian satu dan berbilang kelas kerja dapat dilakukan tanpa perlu membina dan menyelesaikan model-model baru. Keputusan-keputusan yang meyakinkan bagi model-model peruntukan beban kerja dengan taburan servis yang berbeza memotivasikan usaha untuk memperolehi saiz penimbal yang berhubung secara terus dengan beban kerja yang diberikan dalam suatu rangkaian komputer-komputer. Ungkapan gelung tertutup bagi saiz penimbal untuk satu kelas kerja dan saiz penimbal separa bagi berbilang kelas kerja diterbitkan dan menunjukkan keberhubungan dengan ketibaan beban kerja dan kadar pemprosesan, dengan cara pengiraan yang cekap.

ABSTRACT

Multiple Queue Multiple Server Queueing models are used to model workload allocation problems in a network of computers. The problem of determining optimal allocation of workload with single and multi class jobs to a parallel of computers is presented followed by a problem of determining optimal buffer size related to arrival of workload to a single computer. The generalized exponential (GE) distributional model with known first two moments has been used to represent general inter arrival and service time distributions as various jobs have various traffic characteristic. Other service distributional models such as exponential, Erlang-*k* and Gamma have also been used to expand the work applicability. A new algorithm of workload allocation scheme using First Come First Serve discipline in conjunction with optimization of GE queueing systems is proposed for minimizing mean queue length and mean response time in a network of computer systems. This has an advantage over a classical queueing allocation scheme, and is favorably compared. The performance measures, mean queue length and mean response time of the proposed scheme have practically shown improvement.

The principle of optimization and GE distributional model are used to derive a new workload allocation model of single class jobs in a network of queueing system. The reusable concept is proposed to gain solution for determining individual job allocation in a multi class environment. This study is of interest whereby both single and multi class assumption can be done without repeatedly developing and solving new models.

The convincing results of workload allocation models proposed has motivated the work to obtain the direct dependence of the buffer size on the given workload in the network of computer systems. A closed loop expression for buffer sizing of single class jobs and partial buffer sizing of multi class jobs are derived and show their dependency on workload arrival and processing rate in a computationally efficient way.

ACKNOWLEDGEMENTS

First and for most, I would like to thank God for giving me good health, courage and patience, which have enabled me to pursue this doctoral study.

I would like to thank my research supervisor, Professor Dr. Hajah Ku Ruhana Ku Mahamud, for her help and guidance throughout the research process. Without her careful supervision and expertise, this thesis would not have been possible. She has been a valuable mentor, and has helped me to mature as a researcher and a scholar.

I am also indebted to the Government of Malaysia (Ministry of Technology and Environmental) and Universiti Utara Malaysia for sponsoring this research.

I would like to thank my family. I could not have made it to this point without the love and support from them. My parents, they have always believed in me and have always encouraged me to pursue my dreams.

Last but not least, I would like to thank my husband, Issham and my lovely daughter, Siti Zubaidah. Without their love, patience, sense of humor, and understanding, none of this would have been possible.

TABLE OF CONTENTS

		Page
PERMISSION TO USE		ii
ABSTRAK		iii
ABSTRACT		v
ACKNOWLEDGMENTS		vii
LIST OF TABLES		xii
LIST OF FIGURES		xxiv
LIST OF ABBREVIATIONS		xxii
CHAPTER ONE : INTRODUCTION		
1.1	Background	1
1.2	Problem Statement	7
1.3	Objective of the Research	10
1.4	Significance of the Research	12
1.5	Method of the Research	12
1.6	Scope and Assumption of the Research	13
1.7	Organization of the Thesis	13
CHAPTER TWO : LITERATURE REVIEW		
2.1	Introduction	16
2.2	Workload Model	17
	2.2.1 Single Class Job	18
	2.2.2 Multi Class Job	19
2.3	Workload Allocation in A Network of Computers	20
	2.3.1 Terminology	20
	2.3.2 Workload Allocation Taxonomy	22
	2.3.3 Workload Allocation Objective and Performance Measure	23
	2.3.4 Allocation Mechanism	24
2.4	Workload Allocation Policy	25
	2.4.1 Static Policy	25
	2.4.2 Dynamic Policy	28
	2.4.3 Comparative Study of Static and Dynamic Policy	30
2.5	Optimization of Queueing System	31
	2.5.1 Queueing Theory	31
	2.5.2 Optimization	34

2.5.2.1 Mathematical Programming 35
2.5.2.2 Graph Model 35
2.5.2.3 Local Search 36
2.5.2.4 Branch-and-Bound 36
2.5.2.5 Dynamic Optimization 37
2.6 Recent Work on Workload Allocation with Single and Multi Class Job 37
2.6.1 Workload Allocation with Single Class Job 37
2.6.2 Workload Allocation with Multi Class Job 40
2.7 Conclusions 42

CHAPTER THREE : OPTIMIZATION OF GENERALIZED EXPONENTIAL MODEL FOR WORKLOAD ALLOCATION

3.1 Introduction 45
3.2 Motivation for Optimization 45
3.3 Generalized Exponential (GE) Distributional Model 46
3.4 Diffusion Approximation 49
3.4.1 Fluid Flow approximation 49
3.4.2 Diffusion Process 51
3.5 Optimized GE Workload Allocation Scheme 55
3.6 Conclusions 59

CHAPTER FOUR : OPTIMAL WORKLOAD ALLOCATION IN A NETWORK OF COMPUTERS WITH SINGLE CLASS JOB

4.1 Introduction 60
4.2 Optimal Workload Allocation in a Network of Computers with General Exponential (GE) Interarrival and Service Distribution 61
4.2.1 Mathematical Model Description 62
4.2.2 Computational Result 65
4.2.3 Model Reduction to Exponential Interarrival and Service time 74
4.2.4 Computational Result 76
4.2.5 Extension to Other Service Times Distribution 85
4.2.6 Computational Result 86
4.3 Model Development 103
4.4 Conclusions 104

CHAPTER FIVE : WORKLOAD ALLOCATION MODEL IN MULTI CLASS JOB ENVIRONMENT

5.1 Introduction 107
5.2 Model Description 109
5.3 Computational Result and Sensitivity Analysis 112
5.3.1 Workload Allocation of Two Class Job

in 2-GE/GE/1 System 113
5.3.2 Workload Allocation of Two Class Job in 2-M/M/1 System 121
Workload Allocation of Two Class Job in 2-M/Erlang-*k*/1 System 130
5.3.4 Workload Allocation of Two Class Job in 2-M/Gamma/1 System 138
5.4 Model Validation 147
5.5 Conclusions 148

CHAPTER SIX : BUFFER SIZING TO MINIMIZE RESPONSE TIME IN A SINGLE CLASS JOB SINGLE COMPUTER SYSTEM

6.1 Introduction 149
6.2 Single Server-Buffer Sizing Model 151
6.3 Buffer Sizing Optimization Model 152
6.3.1 Numerical Example of GE/GE/1 Buffer Sizing Model 154
6.3.2 Simulation Result 160
6.4 Exponential Interarrival and Exponential/Non-Exponential Service Process 161
6.4.1 Numerical Example of M/M/1 Buffer Sizing Model 164
6.4.2 Simulation Result 166
6.4.3 Numerical Example of M/Erlang-*k*/1 and M/Gamma/1 Buffer Sizing Model 168
6.4.4 Simulation Result 175
6.5 Conclusions 179

CHAPTER SEVEN : PARTIAL BUFFER SPACE ALLOCATION TO MINIMIZE RESPONSE TIME IN A MULTI CLASS JOB ENVIRONMENT

7.1 Introduction 180
7.2 Single Server-Partial Buffer Sharing Model 181
7.3 Partial Buffer Sizing Optimization Model 182
7.3.1 GE/GE/1 Partial Buffer Sizing Model 182
7.3.2 Numerical Example of GE/GE/1 Partial Buffer Sizing Model 184
7.3.3 Partial Buffer Sizing with Poisson Arrival and Exponential/Non-Exponential Service Process 187
7.3.4 Numerical Example of M/M/1 Partial Buffer Sizing Model 188
7.3.5 Partial Buffer Sizing with Erlang-*k* and Gamma Service Process 191
7.3.6 Numerical Example of M/Erlang-*k*/1 and M/Gamma/1 Partial Buffer Sizing Model 192
7.4 Conclusions 201

CHAPTER EIGHT : CONCLUSION

8.1 Summary of Chapters 202
8.2 Summary of the GE Workload Allocation Approaches 204
8.3 Research Contribution 205
8.4 Limitation and Suggestion for Further Research 208

BIBLIOGRAPHY 211

APPENDIX A: Mean Queue Length for GE/GE/1 model 223

APPENDIX B: Proof of Theorem 3.2 224

APPENDIX C: Workload Allocation with Generalised Exponential (N-GE/GE/1) 227

APPENDIX D: Workload Allocation with Exponential Service Distribution (N-M/M/1) 230

APPENDIX E: Workload Allocation with Erlang/Gamma Service Distribution (*N*-M/Erlang-*k*/1 or *N*-M/Gamma/1) 233

APPENDIX F: A Closed Loop Expression for Buffer Sizing of Single Class Jobs 236

APPENDIX G: A Closed Loop Expression for Buffer Sizing of Multi Class Jobs 238

APPENDIX H: A Closed Loop Expression for Buffer Sizing of M/Gamma/1 and M/Erlang-*k*/1 multi class jobs 241

APPENDIX I: Simulation Model of Single Class Workload Allocation Model 245

APPENDIX J: Simulation Model of Multi Class Workload Allocation Model 246

VITAE

LIST OF TABLES

Table 4.1: Results for the classical and proposed approaches of a dual GE/GE/1 with $Ca_1^2 = 0.5$, $Ca_2^2 = 0.3$, $Cs_1^2 = 0.2$, $Cs_2^2 = 0.4$ and $\mu_1 = 3$, $\mu_2 = 4$. 65

Table 4.2: Results for the classical and proposed approaches of a dual GE/GE/1 with $Ca_1^2 = 0.1$, $Ca_2^2 = 0.2$, $Cs_1^2 = 0.4$, $Cs_2^2 = 0.3$ and $\mu_1 = 3$, $\mu_2 = 4$. 68

Table 4.3: Results for the classical and proposed approaches of a dual M/M/1 with $\mu_1 = 2$ and $\mu_2 = 1$. 76

Table 4.4: Results for the classical and proposed approaches of a dual M/M/1 with $\mu_1 = 3$ and $\mu_2 = 4$. 79

Table 4.5: Results for the classical and proposed approaches of a dual M/Erlang-k/1 with $k_1 = 2$, $k_2 = 2$ and $\mu_1 = 3$, $\mu_2 = 4$. 86

Table 4.6: Results for the classical and proposed approaches of a dual M/Erlang-k/1 with $k_1 = 2$, $k_2 = 3$ and $\mu_1 = 3$, $\mu_2 = 4$. 89

Table 4.7: Results for the classical and proposed approaches of a dual M/Erlang-k/1 with $k_1 = 3$, $k_2 = 2$ and $\mu_1 = 3$, $\mu_2 = 4$. 92

Table 4.8: Results for the classical and proposed approaches of a dual M/Gamma/1 with $Cs_1^2 = 2$, $Cs_2^2 = 3$ and $\mu_1 = 3$, $\mu_2 = 4$. 95

Table 4.9: Results for the classical and proposed approaches of a dual M/Gamma/1 with $Cs_1^2 = 5$, $Cs_2^2 = 2$ and $\mu_1 = 3$, $\mu_2 = 4$. 98

Table 4.10: Results for the classical and proposed approaches of a dual M/Gamma/1 with $Cs_1^2 = 0.1$, $Cs_2^2 = 0.3$ and $\mu_1 = 3$, $\mu_2 = 4$. 101

Table 5.1: Results for the proposed approach of a two class job of 2-GE/GE/1 with $d_1 = 0.6$, $d_2 = 0.4$, $Ca_1^2 = 0.8$, $Ca_2^2 = 0.3$, $Cs_1^2 = 0.2$, $Cs_2^2 = 0.4$ and $\mu_{11} = 1$, $\mu_{21} = 2$, $\mu_{12} = 1.5$, $\mu_{22} = 2.5$ 113

Table 5.2: Results for the proposed approach of a two class job of 2-GE/GE/1 with $d_1 = 0.6$, $d_2 = 0.4$, $Ca_1^2 = 0.4$, $Ca_2^2 = 0.3$, $Cs_1^2 = 0.2$, $Cs_2^2 = 0.4$ and $\mu_{11} = 1$, $\mu_{21} = 2$, $\mu_{12} = 1.5$, $\mu_{22} = 2.5$ 115

Table 5.3: Results for the proposed approach of a two class job of 2-GE/GE/1 with $d_1 = 0.6$, $d_2 = 0.4$, $Ca_1^2 = 0.3$, $Ca_2^2 = 0.8$, $Cs_1^2 = 0.2$, $Cs_2^2 = 0.4$ and $\mu_{11} = 1$, $\mu_{21} = 2$, $\mu_{12} = 1.5$, $\mu_{22} = 2.5$ 116

Table 5.4: Results for the proposed approach of a two class job of 2-GE/GE/1 with $d_1 = 0.6$, $d_2 = 0.4$, $Ca_1^2 = 0.2$, $Ca_2^2 = 0.4$, $Cs_1^2 = 0.8$, $Cs_2^2 = 0.3$ and $\mu_{11} = 1$, $\mu_{21} = 2$, $\mu_{12} = 1.5$, $\mu_{22} = 2.5$ 118

Table 5.5: Results for the proposed approach of a two class job of 2-GE/GE/1 with $d_1 = 0.6$, $d_2 = 0.4$, $Ca_1^2 = 0.2$, $Ca_2^2 = 0.4$, $Cs_1^2 = 0.8$, $Cs_2^2 = 0.3$ and $\mu_{11} = 1$, $\mu_{21} = 2$, $\mu_{12} = 1.5$, $\mu_{22} = 2.5$ 119

Table 5.6: Results for the proposed approach of a two class job of 2-GE/GE/1 with $d_1 = 0.6$, $d_2 = 0.4$, $Ca_1^2 = 0.2$, $Ca_2^2 = 0.4$, $Cs_1^2 = 0.3$, $Cs_2^2 = 0.8$ and $\mu_{11} = 1$, $\mu_{21} = 2$, $\mu_{12} = 1.5$, $\mu_{22} = 2.5$ 120

Table 5.7: Results for the proposed approach of a two class job of 2-M/M/1 with $d_1 = 0.6$, $d_2 = 0.4$, $\mu_{11} = 1$, $\mu_{21} = 1$, $\mu_{12} = 0.5$, $\mu_{22} = 0.5$ 122

Table 5.8: Results for the proposed approach of a two class job of 2-M/M/1 with $d_1 = 0.6$, $d_2 = 0.4$, $\mu_{11} = 0.8$, $\mu_{21} = 1.2$, $\mu_{12} = 0.4$, $\mu_{22} = 0.6$ 123

Table 5.9: Results for the proposed approach of a two class job of 2-M/M/1 with $d_1 = 0.3$, $d_2 = 0.7$, $\mu_{11} = 1$, $\mu_{21} = 2$, $\mu_{12} = 1$, $\mu_{22} = 3$ 125

Table 5.10: Results for the proposed approach of a two class job of 2-M/M/1 with $d_1 = 0.3$, $d_2 = 0.7$, $\mu_{11} = 2$, $\mu_{21} = 1$, $\mu_{12} = 2$, $\mu_{22} = 2$ 126

Table 5.11: Results for the proposed approach of a two class job of 2-M/M/1 with $d_1 = 0.3$, $d_2 = 0.7$, $\mu_{11} = 1$, $\mu_{21} = 2$, $\mu_{12} = 1$, $\mu_{22} = 3$ 127

Table 5.12: Results for the proposed approach of a two class job of 2-M/M/1 with $d_1 = 0.3$, $d_2 = 0.7$, $\mu_{11} = 2$, $\mu_{21} = 1$, $\mu_{12} = 2$, $\mu_{22} = 2$ 129

Table 5.13: Results for the proposed approach of a two class job of 2-M/M/1 with $d_1 = 0.4$, $d_2 = 0.6$, $k_1 = 2$, $k_2 = 2$, $\mu_{11} = 1$, $\mu_{21} = 2$, $\mu_{12} = 1$, $\mu_{22} = 3$ 130

Table 5.14: Results for the proposed approach of a two class job of 2-M/Erlang-k/1 with $d_1 = 0.4$, $d_2 = 0.6$, $k_1 = 2$, $k_2 = 2$, $\mu_{11} = 2$, $\mu_{21} = 1$, $\mu_{12} = 2$, $\mu_{22} = 2$ 132

Table 5.15: Results for the proposed approach of a two class job of 2-M/Erlang-k/1 with $d_1 = 0.6$, $d_2 = 0.4$, $k_1 = 2$, $k_2 = 3$, $\mu_{11} = 1$, $\mu_{21} = 2$, $\mu_{12} = 1$, $\mu_{22} = 3$ 133

Table 5.16: Results for the proposed approach of a two class job of 2-M/Erlang-k/1 with $d_1 = 0.6$, $d_2 = 0.4$, $k_1 = 2$, $k_2 = 3$, $\mu_{11} = 2$, $\mu_{21} = 1$, $\mu_{12} = 2$, $\mu_{22} = 2$ 134

Table 5.17: Results for the proposed approach of a two class job of 2-M/Erlang-k/1 with $d_1 = 0.6$, $d_2 = 0.4$, $k_1 = 3$, $k_2 = 2$, $\mu_{11} = 1$, $\mu_{21} = 2$, $\mu_{12} = 1$, $\mu_{22} = 3$ 136

Table 5.18: Results for the proposed approach of a two class job of 2-M/Erlang-k/1 with $d_1 = 0.6$, $d_2 = 0.4$, $k_1 = 3$, $k_2 = 2$, $\mu_{11} = 2$, $\mu_{21} = 1$, $\mu_{12} = 2$, $\mu_{22} = 2$ 137

Table 5.19: Results for the proposed approach of a two class job of 2-M/Gamma/1 with $d_1 = 0.6$, $d_2 = 0.4$, $Cs_1^2 = 2$, $C_{s2}^2 = 3$, $\mu_{11} = 1$, $\mu_{21} = 2$, $\mu_{12} = 1$, $\mu_{22} = 3$ 139

Table 5.20: Results for the proposed approach of a two class job of 2-M/Gamma/1 with $d_1 = 0.6$, $d_2 = 0.4$, $Cs_1^2 = 2$, $C_{s2}^2 = 3$, $\mu_{11} = 2$, $\mu_{21} = 1$, $\mu_{12} = 2$, $\mu_{22} = 2$ 140

Table 5.21: Results for the proposed approach of a two class job of 2-M/Gamma/1 with $d_1 = 0.6$, $d_2 = 0.4$, $Cs_1^2 = 5$, $C_{s2}^2 = 2$, $\mu_{11} = 1$, $\mu_{21} = 2$, $\mu_{12} = 1$, $\mu_{22} = 3$ 142

Table 5.22: Results for the proposed approach of a two class job of 2-M/Gamma/1 with $d_1 = 0.6$, $d_2 = 0.4$, $Cs_1^2 = 5$, $C_{s2}^2 = 2$, $\mu_{11} = 2$, $\mu_{21} = 1$, $\mu_{12} = 2$, $\mu_{22} = 2$ 143

Table 5.23: Results for the proposed approach of a two class job of 2-M/Gamma/1 with $d_1 = 0.6$, $d_2 = 0.4$, $Cs_1^2 = 0.1$, $C_{s2}^2 = 0.3$, $\mu_{11} = 1$, $\mu_{21} = 2$, $\mu_{12} = 1$, $\mu_{22} = 3$ 145

Table 5.24: Results for the proposed approach of a two class job of 2-M/Gamma/1 with $d_1 = 0.6$, $d_2 = 0.4$, $Cs_1^2 = 0.1$, $C_{s2}^2 = 0.3$, $\mu_{11} = 2$, $\mu_{21} = 1$, $\mu_{12} = 2$, $\mu_{22} = 2$ 146

Table 6.1: Proposed model mean response time: *PR* versus Simulation model mean response time: *SR* for GE/GE/1 buffer sizing with Ca^2 =0.2, C_s^2 = 0.1, t = 0.05 and λ= 1, 1.2, 1.4,..., 2.0 160

Table 6.2: Proposed model mean response time: *PR* versus Simulation model mean response time: *SR* for GE/GE/1 buffer sizing with Ca^2 =0.3, C_s^2 = 0.6, t = 0.05 and λ= 1, 1.2, 1.4,..., 2.0 161

Table 6.3: Proposed model mean response time: *PR* versus Simulation model mean response time: *SR* for GE/GE/1 buffer sizing with Ca^2 =0.8, C_s^2 = 0.6, t = 0.05 and λ= 1, 1.2, 1.4,..., 2.0 166

Table 6.4: Proposed model mean response time: *PR* versus Simulation model mean response time: *SR* for M/M/1 buffer sizing with t = 0.05 and λ= 2.2, 2.4, 2.6,..., 3.2 176

Table 6.5: Proposed model mean response time: *PR* versus Simulation model mean response time: *SR* for M/Erlang-*2*/1 buffer sizing with t = 0.05 and λ= 1, 1.2, 1.4,..., 2.0 176

Table 6.6: Proposed model mean response time: *PR* versus Simulation model mean response time: *SR* for M/Erlang-*3*/1 buffer sizing with t = 0.05 and λ= 1, 1.2, 1.4,..., 2.0 177

Table 6.7: Proposed model mean response time: *PR* versus Simulation model mean response time: *SR* for M/Erlang-*4*/1 buffer sizing with t = 0.05 and λ= 1, 1.2, 1.4,..., 2.0 177

Table 6.8: Proposed model mean response time: *PR* versus Simulation model mean response time: *SR* for M/Gamma/1 buffer sizing with Cs^2 = 0.05, t = 0.05 and λ= 1, 1.2, 1.4,..., 2.0 178

Table 6.9: Proposed model mean response time: *PR* versus Simulation model mean response time: *SR* for M/Gamma/1 buffer sizing with Cs^2 = 0.5, t = 0.05 and λ= 1, 1.2, 1.4,..., 2.0 178

Table 6.10: Proposed model mean response time: *PR* versus Simulation model mean response time: *SR* for M/Gamma/1 buffer sizing with Cs^2 = 2.5, t = 0.05 and λ= 1, 1.2, 1.4,..., 2.0 178

Table 7.1: Partial buffer size results for the classical and proposed approaches of a GE/GE/1 two class job with $Ca^2 = 0.5$, $Cs^2 = 0.2$, $\lambda_1 = 1$, $\lambda_2 = 2$ and $P_1 = 3$, $P_2 = 4$. 184

Table 7.2: Partial buffer size results for the classical and proposed approaches of a GE/GE/1 two class job with $Ca^2 = 0.2$, $Cs^2 = 0.4$, $\lambda_1 = 2$, $\lambda_2 = 3$ and $P_1 = 3$, $P_2 = 4$. 185

Table 7.3: Partial buffer size results for the classical and proposed approaches of a M/M/1 two class job with , $\lambda_1 = 1$, $\lambda_2 = 2$ and $P_1 = 3$, $P_2 = 4$. 188

Table 7.4: Partial buffer size results for the classical and proposed approaches of a M/M/1 two class job with $\lambda_1 = 2$, $\lambda_2 = 3$ and $P_1 = 3$, $P_2 = 4$. 189

Table 7.5: Partial buffer size results for the classical and proposed approaches of a M/Erlang-2/1 two class job with, $\lambda_1 = 1$, $\lambda_2 = 2$ and $P_1 = 3$, $P_2 = 4$. 192

Table 7.6: Partial buffer size results for the classical and proposed approaches of a M/Erlang-2/1 two class job with $\lambda_1 = 2$, $\lambda_2 = 3$ and $P_1 = 3$, $P_2 = 4$. 193

Table 7.7: Partial buffer size results for the classical and proposed approaches of a M/Erlang-3/1 two class job with , $\lambda_1 = 2$, $\lambda_2 = 3$ and $P_1 = 3$, $P_2 = 4$. 194

Table 7.8: Partial buffer size results for the classical and proposed approaches of a M/Gamma/1 two class job with $c_s^2 = 0.2$, $\lambda_1 = 1$, $\lambda_2 = 2$ and $P_1 = 3$, $P_2 = 4$. 195

Table 7.9: Partial buffer size results for the classical and proposed approaches of a M/Gamma/1 two class job with $c_s^2 = 0.2$, $\lambda_1 = 2$, $\lambda_2 = 3$ and $P_1 = 3$, $P_2 = 4$. 196

Table 7.10: Partial buffer size results for the classical and proposed approaches of a M/Gamma/1 two class job with $c_s^2 = 0.8$, $\lambda_1 = 1$, $\lambda_2 = 2$ and $P_1 = 3$, $P_2 = 4$. 198

Table 7.11: Partial buffer size results for the classical and proposed approaches of a M/Gamma/1 two class job with $c_s^2 = 0.8$, $\lambda_1 = 2$, $\lambda_2 = 3$ and $P_1 = 3$, $P_2 = 4$. 199

LIST OF FIGURES

Figure 1.1: A central job routing system 4

Figure 1.2: Multiple Queue Multiple Server Model 5

Figure 2.1: Queueing-based taxonomy of workload allocation 23

Figure 2.2: Single server queueing model 32

Figure 2.3: Classification of earlier work on static workload allocation using queueing based taxonomy 42

Figure 4.1: Performance improvement of a dual GE/GE/1 with $Ca_1^2 = 0.5$, $Ca_2^2 = 0.3$, $Cs_1^2 = 0.2$, $Cs_2^2 = 0.4$ and $\mu_1 = 3$, $\mu_2 = 4$ 66

Figure 4.2: Analytical versus simulation result for a a dual GE/GE/1 with $Ca_1^2 = 0.5$, $Ca_2^2 = 0.3$, $Cs_1^2 = 0.2$, $Cs_2^2 = 0.4$ and $\mu_1 = 3$, $\mu_2 = 4$ 67

Figure 4.3: Performance improvement of a dual GE/GE/1 with $Ca_1^2 = 0.1$, $Ca_2^2 = 0.2$, $Cs_1^2 = 0.4$, $Cs_2^2 = 0.3$ and $\mu_1 = 3$, $\mu_2 = 4$ 69

Figure 4.4: Performance improvement of a dual GE/GE/1 with $Ca_1^2 = 0.1$, $Ca_2^2 = 0.2$, $Cs_1^2 = 0.4$, $Cs_2^2 = 0.3$ and $\mu_1 = 3$, $\mu_2 = 4$. 70

Figure 4.5: Performance improvement for a sample number of computers where $\rho = 0.9$ 71

Figure 4.6: Performance improvement of a dual M/M/1 with $\mu_1 = 2$ and $\mu_2 = 1$. 77

Figure 4.7: Analytical versus simulation result for a dual M/M/1 with $\mu_1 = 2$ and $\mu_2 = 1$ 78

Figure 4.8: Performance improvement of a dual M/M/1 with $\mu_1 = 3$ and $\mu_2 = 4$. 80

Figure 4.9: Analytical versus simulation result for a dual M/M/1 with $\mu_1 = 3$ and $\mu_2 = 4$ 81

Figure 4.10: Performance improvement of a dual M/Erlang-k/1 with $k_1 = 2$, $k_2 = 2$ $\mu_1 = 3$ and $\mu_2 = 4$ 87

Figure 4.11: Analytical versus simulation result for a dual M/Erlang-k/1 with $k_1 = 2$, $k_2 = 2$ and $\mu_1 = 3$, $\mu_2 = 4$ 88

Figure 4.12: Performance improvement of a dual M/Erlang-k/1 with $k_1 = 2$, $k_2 = 3$ $\mu_1 = 3$ and $\mu_2 = 4$ 90

Figure 4.13: Analytical versus simulation result for a dual M/Erlang-k/1 with $k_1 = 2$, $k_2 = 3$ and $\mu_1 = 3$, $\mu_2 = 4$ 91

Figure 4.14: Performance improvement of a dual M/Erlang-k/1 with $k_1 = 3$, $k_2 = 2$ $\mu_1 = 3$ and $\mu_2 = 4$ 93

Figure 4.15: Analytical versus simulation result for a dual M/Erlang-k/1 with $k_1 = 3$, $k_2 = 2$ and $\mu_1 = 3$, $\mu_2 = 4$ 94

Figure 4.16: Performance improvement of a dual M/Gamma/1 with $Cs_1^2 = 2$, $Cs_2^2 = 3$ and $\mu_1 = 3$, $\mu_2 = 4$ 96

Figure 4.17: Analytical versus simulation result for a dual M/Gamma/1 with $Cs_1^2 = 2$, $Cs_2^2 = 3$ and $\mu_1 = 3$, $\mu_2 = 4$ 97

Figure 4.18: Performance improvement of a dual M/Gamma/1 with $Cs_1^2 = 5$, $Cs_2^2 = 2$ and $\mu_1 = 3$, $\mu_2 = 4$ 99

Figure 4.19: Performance improvement of a dual M/Gamma/1 with $Cs_1^2 = 5$, $Cs_2^2 = 2$ and $\mu_1 = 3$, $\mu_2 = 4$ 100

Figure 4.20: Performance improvement of a dual M/Gamma/1 with $Cs_1^2 = 0.1$, $Cs_2^2 = 0.3$ and $\mu_1 = 3$, $\mu_2 = 4$ 102

Figure 4.21: Performance improvement of a dual M/Gamma/1 with $Cs_1^2 = 0.1$, $Cs_2^2 = 0.3$ and $\mu_1 = 3$, $\mu_2 = 4$ 103

Figure 5.1: The decomposition model of m-class jobs to a network of n-computers 108

Figure 5.2: Model of two-class job to a network of *2*-computers 110

Figure 5.3: Analytical versus simulation result for a two-class of 2-GE/GE/1 with $d_1 = 0.6$, $d_2 = 0.4$, $Ca_1^2 = 0.8$, $Ca_2^2 = 0.3$, $Cs_1^2 = 0.2$, $Cs_2^2 = 0.4$ and $\mu_{11} = 1$, $\mu_{21} = 2$, $\mu_{12} = 1.5$, $\mu_{22} = 2.5$ 114

Figure 5.4: Analytical versus simulation result for a two-class of 2-GE/GE/1 with $d_1 = 0.6$, $d_2 = 0.4$, $Ca_1^2 = 0.4$, $Ca_2^2 = 0.3$, $Cs_1^2 = 0.2$, $Cs_2^2 = 0.4$ and $\mu_{11} = 1$, $\mu_{21} = 2$, $\mu_{12} = 1.5$, $\mu_{22} = 2.5$ 116

Figure 5.5: Analytical versus simulation result for a two-class of 2-GE/GE/1 with $d_1 = 0.6$, $d_2 = 0.4$, $Ca_1^2 = 0.3$, $Ca_2^2 = 0.8$, $Cs_1^2 = 0.2$, $Cs_2^2 = 0.4$ and $\mu_{11} = 1$, $\mu_{21} = 2$, $\mu_{12} = 1.5$, $\mu_{22} = 2.5$ 117

Figure 5.6: Analytical versus simulation result for a two-class of 2-GE/GE/1 with $d_1 = 0.6$, $d_2 = 0.4$, $Ca_1^2 = 0.2$, $Ca_2^2 = 0.4$, $Cs_1^2 = 0.8$, $Cs_2^2 = 0.3$ and $\mu_{11} = 1.5$, $\mu_{21} = 1.5$, $\mu_{12} = 2$, $\mu_{22} = 2$ 118

Figure 5.7: Analytical versus simulation result for a two-class of 2-GE/GE/1 with $d_1 = 0.6$, $d_2 = 0.4$, $Ca_1^2 = 0.2$, $Ca_2^2 = 0.4$, $Cs_1^2 = 0.8$, $Cs_2^2 = 0.3$ and $\mu_{11} = 1$, $\mu_{21} = 2$, $\mu_{12} = 1.5$, $\mu_{22} = 2.5$ 120

Figure 5.8: Analytical versus simulation result for a two-class of 2-GE/GE/1 with $d_1 = 0.6$, $d_2 = 0.4$, $Ca_1^2 = 0.2$, $Ca_2^2 = 0.4$, $Cs_1^2 = 0.8$, $Cs_2^2 = 0.3$ and $\mu_{11} = 1$, $\mu_{21} = 2$, $\mu_{12} = 1.5$, $\mu_{22} = 2.5$ 121

Figure 5.9: Analytical versus simulation result for a two-class of 2-M/M/1 with $d_1 = 0.6$, $d_2 = 0.4$, $\mu_{11} = 1$, $\mu_{21} = 1$, $\mu_{12} = 0.5$, $\mu_{22} = 0.5$ 123

Figure 5.10: Analytical versus simulation result for a two-class of 2-M/M/1 with $d_1 = 0.6$, $d_2 = 0.4$, $\mu_{11} = 0.8$, $\mu_{21} = 1.2$, $\mu_{12} = 0.4$, $\mu_{22} = 0.6$ 124

Figure 5.11: Analytical versus simulation result for a two-class of 2-M/M/1 with $d_1 = 0.3$, $d_2 = 0.7$, $\mu_{11} = 1$, $\mu_{21} = 2$, $\mu_{12} = 1$, $\mu_{22} = 3$ 125

Figure 5.12: Analytical versus simulation result for a two-class of 2-M/M/1 with $d_1 = 0.3$, $d_2 = 0.7$, $\mu_{11} = 2$, $\mu_{21} = 1$, $\mu_{12} = 2$, $\mu_{22} = 2$ 127

Figure 5.13: Analytical versus simulation result for a two-class of 2-M/M/1 with $d_1 = 0.3$, $d_2 = 0.7$, $\mu_{11} = 1$, $\mu_{21} = 2$, $\mu_{12} = 1$, $\mu_{22} = 3$ 128

Figure 5.14: Analytical versus simulation result for a two-class of 2-M/M/1 with $d_1 = 0.3$, $d_2 = 0.7$, $\mu_{11} = 2$, $\mu_{21} = 1$, $\mu_{12} = 2$, $\mu_{22} = 2$ 129

Figure 5.15: Analytical versus simulation result for a two-class of 2-M/Erlang-k/1 with $d_1 = 0.4$, $d_2 = 0.6$, $k_1 = 2$, $k_2 = 2$, $\mu_{11} = 1$, $\mu_{21} = 2$, $\mu_{12} = 1$, $\mu_{22} = 3$ 131

Figure 5.16: Analytical versus simulation result for a two-class of 2-M/Erlang-k/1 with $d_1 = 0.4$, $d_2 = 0.6$, $k_1 = 2$, $k_2 = 2$, $\mu_{11} = 2$, $\mu_{21} = 1$, $\mu_{12} = 2$, $\mu_{22} = 2$ 132

Figure 5.17: Analytical versus simulation result for a two-class of 2-M/Erlang-k/1 with $d_1 = 0.6$, $d_2 = 0.4$, $k_1 = 2$, $k_2 = 3$, $\mu_{11} = 1$, $\mu_{21} = 2$, $\mu_{12} = 1$, $\mu_{22} = 3$ 134

Figure 5.18: Analytical versus simulation result for a two-class of 2-M/Erlang-k/1 with $d_1 = 0.6$, $d_2 = 0.4$, $k_1 = 2$, $k_2 = 3$, $\mu_{11} = 2$, $\mu_{21} = 1$, $\mu_{12} = 2$, $\mu_{22} = 2$ 135

Figure 5.19: Analytical versus simulation result for a two-class of 2-M/Erlang-k/1 with $d_1 = 0.6$, $d_2 = 0.4$, $k_1 = 3$, $k_2 = 2$, $\mu_{11} = 1$, $\mu_{21} = 2$, $\mu_{12} = 1$, $\mu_{22} = 3$ 136

Figure 5.20: Analytical versus simulation result for a two-class of 2-M/Erlang-k/1 with $d_1 = 0.6$, $d_2 = 0.4$, $k_1 = 3$, $k_2 = 2$, $\mu_{11} = 2$, $\mu_{21} = 1$, $\mu_{12} = 2$, $\mu_{22} = 2$ 138

Figure 5.21: Analytical versus simulation result for a two-class of 2-M/Gamma/1 with $d_1 = 0.6$, $d_2 = 0.4$, $Cs_1^2 = 2$, $C_{s2}^2 = 3$, $\mu_{11} = 1$, $\mu_{21} = 2$, $\mu_{12} = 1$, $\mu_{22} = 3$ 140

Figure 5.22: Analytical versus simulation result for a two-class of 2-M/Gamma/1 with $d_1 = 0.6$, $d_2 = 0.4$, $Cs_1^2 = 2$, $C_{s2}^2 = 3$, $\mu_{11} = 2$, $\mu_{21} = 1$, $\mu_{12} = 2$, $\mu_{22} = 2$ 141

Figure 5.23: Analytical versus simulation result for a two-class of 2-M/Gamma/1 with $d_1 = 0.6$, $d_2 = 0.4$, $Cs_1^2 = 5$, $C_{s2}^2 = 2$, $\mu_{11} = 1$, $\mu_{21} = 2$, $\mu_{12} = 1$, $\mu_{22} = 3$ 143

Figure 5.24: Analytical versus simulation result for a two-class of 2-M/Gamma/1 with $d_1 = 0.6$, $d_2 = 0.4$, $Cs_1^2 = 5$, $C_{s2}^2 = 2$, $\mu_{11} = 2$, $\mu_{21} = 1$, $\mu_{12} = 2$, $\mu_{22} = 2$ 144

Figure 5.25: Analytical versus simulation result for a two-class of 2-M/Gamma/1 with $d_1 = 0.6$, $d_2 = 0.4$, $Cs_1^2 = 0.1$, $C_{s2}^2 = 0.3$, $\mu_{11} = 1$, $\mu_{21} = 2$, $\mu_{12} = 1$, $\mu_{22} = 3$ 145

Figure 5.26: Analytical versus simulation result of a two-class job of 2- M/Gamma/1 with $d_1 = 0.6$, $d_2 = 0.4$, $Cs_1^2 = 0.1$, $Cs_2^2 = 0.3$, $\mu_{11} = 2$, $\mu_{21} = 1$, $\mu_{12} = 2$, $\mu_{22} = 2$. 147

Figure 6.1: A network link modeled by a single server finite-buffer system 151

Figure 6.2: Buffer allocation of single class job of GE/GE/1 with Ca^2 =0.2, C_s^2 = 0.1, t = 0.05 and λ = 1, 1.2, 1.4,..., 2.0 156

Figure 6.3: Buffer size-Processing rate-Response time function of GE/GE/1 with Ca^2 =0.2, C_s^2 = 0.1, t = 0.05 and ρ = 0.9 156

Figure 6.4: Buffer allocation of single class job of GE/GE/1 with Ca^2 =0.3, C_s^2 = 0.6, t = 0.05 and λ = 1, 1.2, 1.4,..., 2.0 157

Figure 6.5: Buffer size-Processing rate-Response time function of GE/GE/1 with Ca^2 =0.3, C_s^2 = 0.6, t = 0.05 and ρ = 0.9 158

Figure 6.6: Buffer allocation of single class job of GE/GE/1 with Ca^2 =0.8, C_s^2 = 0.6, t = 0.05 and λ = 1, 1.2, 1.4,..., 2.0 158

Figure 6.7: Buffer size-Processing rate-Response time function of GE/GE/1 with Ca^2 =0.8, C_s^2 = 0.6, t = 0.05 and ρ = 0.9 159

Figure 6.8: Buffer allocation of single class job M/M/1 with t = 0.05 and λ = 2.2, 2.4, 2.6,..., 3.2 165

Figure 6.9: Buffer size-Processing rate-Response time function of M/M/1 with t = 0.05 and ρ = 0.9 165

Figure 6.10: Buffer allocation of single class job M/Erlang-*2*/1 with t = 0.05 and λ = 1, 1.2, 1.4,..., 2.0 169

Figure 6.11: Buffer size-Processing rate-Response time function of M/Erlang-*2*/1 with t = 0.05 and ρ = 0.9 170

Figure 6.12: Buffer allocation of single class job M/Erlang-*3*/1 with t = 0.05 and λ = 1, 1.2, 1.4,..., 2.0 170

Figure 6.13: Buffer size-Processing rate-Response time function of M/Erlang-*3*/1 with $t = 0.05$ and $\rho = 0.9$ 171

Figure 6.14: Buffer allocation of single class job M/Erlang-*4*/1 with $t = 0.05$ and $\lambda = 1, 1.2, 1.4, \ldots, 2.0$ 171

Figure 6.15: Buffer size-Processing rate-Response time function of M/Erlang-*4*/1 with $t = 0.05$ and $\rho = 0.9$ 172

Figure 6.16: Buffer allocation of single class job M/Gamma/1 with $Ca^2 = 0.05$, $t = 0.05$ and $\lambda = 1, 1.2, 1.4, \ldots, 2.0$ 172

Figure 6.17: Buffer size-Processing rate-Response time function of M/Gamma/1 with $Ca^2 = 0.05$, $t = 0.05$ and $\rho = 0.9$ 173

Figure 6.18: Buffer allocation of single class job M/Gamma/1 with $Ca^2 = 0.5$, $t = 0.05$ and $\lambda = 1, 1.2, 1.4, \ldots, 2.0$ 173

Figure 6.19: Buffer size-Processing rate-Response time function of M/Gamma/1 with $Ca^2 = 0.5$, $t = 0.05$ and $\rho = 0.9$ 174

Figure 6.20: Buffer allocation of single class job M/Gamma/1 with $Ca^2 = 2.5$, $t = 0.05$ and $\lambda = 1, 1.2, 1.4, \ldots, 2.0$ 174

Figure 6.21: Buffer size-Processing rate-Response time function of M/Gamma/1 with $Ca^2 = 2.5$, $t = 0.05$ and $\rho = 0.9$ 175

Figure 7.1: A shared buffer multi class job system with FIFO queue 181

Figure 7.2: Analytical versus simulation result of partial buffer sizing performance of GE/GE/1 two class job with $Ca^2 = 0.5$, $Cs^2 = 0.2$, $\lambda_1 = 1$, $\lambda_2 = 2$ and $P_1 = 3$, $P_2 = 4$. 185

Figure 7.3: Analytical versus simulation result of partial buffer sizing performance of GE/GE/1 two class job with $Ca^2 = 0.2$, $Cs^2 = 0.4$, $\lambda_1 = 2$, $\lambda_2 = 3$ and $P_1 = 3$, $P_2 = 4$. 186

Figure 7.4: Analytical versus simulation result of partial buffer sizing performance of M/M/1 two class job with λ_1 =1, λ_2 =2 and P_1=3, P_2 = 4. 189

Figure 7.5: Analytical versus simulation result of partial buffer sizing performance of M/M/1 two class job with λ_1 = 2, λ_2 =3 and P_1=3, P_2 = 4. 190

Figure 7.6: Analytical versus simulation result of partial buffer sizing performance of M/Erlang-2/1 two class job with λ_1 =1, λ_2 =2 and P_1 = 3, P_2 = 4. 193

Figure 7.7: Analytical versus simulation result of partial buffer sizing performance of M/erlang-2/1two class job with λ_1 = 2, λ_2 =3 and P_1 = 3, P_2 = 4. 194

Figure 7.8: Analytical versus simulation result of partial buffer sizing performance of M/Erlang-3/1 two class job with λ_1 = 2, λ_2 = 3 and P_1 = 3, P_2 = 4. 195

Figure 7.9: Analytical versus simulation result of partial buffer sizing performance of M/Gamma/1 two class job with c_s^2 = 0.2, λ_1 = 1, λ_2 = 2 and P_1 = 3, P_2 = 4. 196

Figure 7.10: Analytical versus simulation result of partial buffer sizing performance of M/Gamma/1 two class job with c_s^2 = 0.2, λ_1 = 2, λ_2 = 3 and P_1 = 3, P_2 = 4. 197

Figure 7.11: Analytical versus simulation result of partial buffer sizing performance of M/Gamma/1 two class job with c_s^2 =0.8, λ_1 = 1, λ_2 = 2 and P_1 = 3, P_2 = 4. 198

Figure 7.12: Analytical versus simulation result of partial buffer sizing performance of M/Gamma/1 two class job with c_s^2 = 0.8, λ_1 = 2, λ_2 = 3 and P_1 = 3, P_2 = 4. 199

LIST OF ABBREVIATIONS

CV	-	coeffiecient of variation
CPU	-	Central Processing Unit
i.i.d	-	independent identically and distributed
I/O	-	input and output
FIFO	-	First in First out
GE	-	generalized exponential
Pdf	-	probability density function
GE/GE/1	-	generalized exponential arrival and service time with single server
M/M/1	-	exponential arrival and service time with single server
M/Erlang-*k*/1	-	exponential arrival and *k* number of service stages with single server
M/Gamma/1	-	exponential arrival and gamma service time with single server

CHAPTER ONE

INTRODUCTION

1.1 Background

"Any successful system must do what its designer wants it to do. If the system cannot meet this basic demand, it is meaningless to talk about any other thing" (Hu & Gorton, 1997). This statement shows how crucial system design is to system development. In this thesis, we are mainly concerned with computer-based systems, for they can process data speedily and accurately.

Computer systems are now becoming part of a larger system such as insurance companies, tax offices, banks, administrative departments and industrial organization. These systems influence daily business operation and also are influenced by the wishes expressed and demands enforced by their environment. In practice this means facing deadlines, dealing with unforeseen circumstances, for example unexpected outcome, delay of work, and the sudden arrival of even more work. These situations commonly involve insufficient resource capacity to serve all jobs completely, and even to grant every single request for service. Hence both time and capacity are precious.

The contents of the thesis is for internal user only

BIBLIOGRAPHY

Allen, A. O. (1990). Probability, *Statistic and Queueing Theory with Computer Science Application*, 2nd. Edition. Academic Press Limited, London.

Albin, S. L. (1982). On Poisson Approximations for Superposition Arrival Processes in Queues. *Management Science*, 28(2), 132-140.

Altman, E., Bhulai, S., Gaujal, B. and Hordijk, A. (1999). Optimal Routing Problems and Multimodularity. *Technical Report INRIA*, Vrije Universiteit, Amsterdam.

Balci, O. (2003). Verification, Validation, and Certification of Modeling and Simulation Applications. In *Proceedings of the 2003 Winter Simulation Conference* (New Orleans, LA, Dec. 7-10). IEEE, Piscataway, NJ, pp. 150-158.

Banawan, S. A. and Zahorjan, J. (1989). Load Sharing in Heterogeneous Queueing Systems. *In: Proceedings IEEE INFOCOM'89*, 731-739.

Banawan, S. A. and Zeidat, N. M. (1992). A Comparative Study of Load Sharing in Heterogeneous Multicomputer Systems. *Proceeding of the 25th Annual Symposium on Simulation*, Orlando, Florida, 22-31.

Bansal, N. and Dhamdhere, K. (2003). Minimizing Weighted ow Time. *In: ACM-SIAM symposium on Discrete Algorithm* (SODA), 508-516.

Baskett, F., Chandy, K. M., Muntz, R. R., and Palacios, F. (1975). Open, Closed and Mixed Networks of Queues with Different Classes of Customers. *Journal of the ACM*, 22(2), 248-260.

Bell, S. L. and Williams, R. J. (1999). Dynamic Scheduling of a System with Two Parallel Servers: Asymptotic optimality of a continuous review threshold policy in heavy Trac. *In: Proceedings of the 38th Conference on Decision and Control*, Pheonix, Arizona, 1743-1748.

Birman, A., Chang, P. C., Chen, J. S.-C., and Guerin, R. (1989). *Buffer Sizing in an ISDN Frame Delay Switch.* Technical Report RC 14386, IBM Research, IBM T. J. Watson Research Center.

Birman, A., Gail, H. R., Hantler, S. L., Rosberg, Z., and Sidi, M. (1991). An Optimal Service Policy for Buffer Systems. *Journal of the ACM*, 42(3): 641 – 657.

Brouns, A.J.F. (2003). Queueing *Model With Admission and Termination Control.* PhD. Thesis, Eindhoven Univerisity. Netherland.

Boel, R.K. and van Schuppen, J.H. (1989). Distributed Routing for Load Balancing. *Proceedings of the IEEE*, 77:210-221.

Bolch, G. (2002). *Performance Modeling of Computer System*, London, McGraw Hill.

Borst, A. C. (1995). Optimal Probabilistic Allocation of Customer Types to Servers. *ACM SIGMETRICS Performance Evaluation Review* , 23(1), 116-125.

Boxma, O. J. (1995). Static Optimization of Queueing Systems. *CWI Report.* BS-R 9302.

Casavant, T. L. and Kuhl, J. G. (1988). A Taxanomy of Sceduling in General-Purpose Distributed Computing Systems. *IEEE Trans. On Software Engineering*, 14(2), 144-154.

Chandy, K. M., Herzog, U. and Woo, L. (1975). Parametric Analysis of Queueing networks. *IBM J. Research and Development*, 19(1), 36-42.

Chanson, S. T., Peng, W., Hui, C. C., Tang, X. and To, M. Y. (1999). Multidomain load Balancing. *Technical Report* HKUST-CS99-18, Department of Computer Science, HKUST.

Chevance, R. J. (2004). *Sever Architectures: Multiprosessors, Clusters, Parallel Systems, Web Servers*, Storage Solutions. Digital Press.

Chombe, M. B. and Boxma, O. J. (1995). Optimization of Static Traffic Allocation Policies. *Theoretical Computer Science*, 125, 17-43.

Chow, Y. C. and Kohler, W. H. (1979). Models of Dynamic Load Balancing in a Heterogeneous Multiple Processor System, *IEEE Transactions on Computers*, C-28(5), 354-361.

De Jongh, J. F. C. M., (2002). Share Scheduling in Distributed System. *PhD Thesis* University of Technische, Netherland.

Dowd, P. and Gelenbe, E. (1995). MASCOTS'95: *Proceedings of the Third International Workshop on Modeling, Analysis and Simulation of Computer and Telecommunication Systems*, IEEE Computer Society.

Eager, D. L., Lazowska, E. D., and Zahorjan, J. (1986a). A Comparison of Receiver-Initiated and Sender-Initiated Adaptive Load Sharing. *Performance Evaluation*, 6, 53-68.

Eager, D. L., Lazowska, E. D., and Zahorjan, J. (1986b). Adaptive Load Sharing in Homogeneous Distributed Systems, *IEEE Transaction on Software Engineering*, 12(5), 662-675.

El-Affendi, M. A. and Koavatsos, D. D. (1983). A Maximum Entrophy Analysis of the M/G/1 and G/M/1 Queueing Systems at Equilibrium, *Acta Informatica*, 19, 339 – 355.

Epema, D. H. J. and de Jong, J. F. C. M. (1999). Proportional Share Scheduling in Single-Server and Multiple-Server Computing Systems. *Performance Evaluation Review*, 27(3), 7-10.

Ephremides, A., Varaiya, P. and Walrand, J. (1980). A Simple Dynamic Routing Problem. *IEEE Trans. Automatic Control.* AC-25, 690-693.

Francois-Dutot, P., Eyraud, L. Mounie, G. and Trystram, D. (2004). Bi-criteria Algorithm for Scheduling jobs on cluster platform. In http://citeseer online.

Franks G., Majumdar, S., Neilson, J., Petriu, D., Rolia, J., Woodside, M. (1996). Performance Analysis of Distributed Server Systems. *6th International Conference on Software Quality (6ICSQ)*, 15-26.

Gail, H. R., Grover, G., Guerin, R., Hantler, S. L., Rosberg, Z. and Sidi, M., (1993). Buffer Size Requirements under Longest Queue First. *Performance Evaluation*, 18(2).

Gelenbe, E and Mitrani, I. (1980). *Analysis and Synthesis of Computer Systems*. London: Academic Press.

Georgiadis, L., Nikolaou, C. and Thomasian, A. (2004). A Fair Workload Allocation Policy for Heterogeneous Systems. *IEEE/ACM Transactions on Networking (TON)*. 1(6), 231-242.

Georgiadis, L., and Guerin, R., (1994), Optimal Multiplexing on a Single Link: Delay and Buffer Requirement. IBM T. J. Watson Research Center, Research Report, RC 19711 (97393).

Ghoneim, H. and Stidham, S. (1985). Control of Arrivals to Two Queues in Series. *European Journal of Operation Research*, 21, 399-409.

Gunther, N. J. (2000). *The Practical Performance Analyst*. Lincoln: McGraw Hill.

Hajek, B. and Seri, P. (2000). Lex-Optimal Multiclass Scheduling with Deadlines. *submitted to Mathematics of Operations Research.*

Harchol-Balter, M., and Downey, A. B., (1997). Exploiting Process Lifetime Distribution for dynamic Load balancing. *ACM Transactions on Computer Systems*, 15(3): 253 – 285.

Harchol-Balter, M., Crovella, M., and Murta, C. (1997). To Queue or Not to Queue: When Queueing is Better Than Timesharing in a Distributed System. *Technical Report BUCS-TR-1997-017.*

Harchol-Balter, M., Crovella, M., and Murta, C. (1999). On Choosing a Task Assignment Policy for a Distributed Server System. *Journal of Parallel Distributed Computing*. 59, 204-228.

Harchol-Balter, M. (2002). Task Assignment with Unknown Duration. *Journal of the ACM*. 49(2).260-288.

Harrell, C. R. and Tumay, K. (1996). *Simulation Made Easy*. A Manager's Guide. Industrial Engineering and Management Press. Georgia.

Harrison, P. G. and Patel, N. M. (1992). *Performance Modeling of Communication Networks and Computer Architecture*. Great Britain: Addison Wesley.

He, L., Jarvis, S. A., Spooner, D. P. and Graham, R. N. (2004). Optimizing Static Workload Allocation in Multiclusters. *IEEE Computers*, 47(2): 287-296

Herschberg, L. S., Epema, D. H. J. and de Jong, J. F. C. M. (1992). *A literature study on Scheduling in Distributed Systems*, Technical Report, University of Delft, Netherland.

Hlavacs, H. and Ueberhuber, C. W. (2001). *Performance Evaluation by Simulation.* Technical Report, UNSW-CSE-TR-9707, University of NSW, Australia.

Hsiao, M. T. and Lazar A. A. (1990). Optimal Flow Control of Multiclass Queueing Networks with Partial Information. *IEEE Transaction on Automatic Control.* 35(7), 855-860.

Hsiao, M. T. and Lazar, A. A. (1991). Optimal Decentralized Flow Control of Markovian Queueing Networks with multiple Controllers. *Performance Evaluation*. 13(3), 181-204.

Hu, L. and Gorton, I. (1997). *Performance Evaluation for Parallel Systems: A Survey*, Technical Report UNSW-CSE-TR-9707, University of NSW, Sydney, Australia, October.

Jain, R. (1991). *The Art of Computer Systems Performance Analysis Techniques for Experimental Design, Measurement, Simulation, and Modeling*, John Wiley & Sons. New York.

Jean-Marie, A. (1987). Load Balancing in a System of Two Queues with Resequencing. *In:Proceeding of Performance '87*, 75-88 (North Holland, Amsterdam, P.J. Courtois, G. Latouche (Eds.)).

Jean-Marie, A and Gun, L. (1993). Parallel Queue with Resequencing. *Journal of the ACM*. 40(5), 1188-1208.

Kameda, H., Said Fathy, El Z., Ryu, I. And Li, J. (1997). A Performance Comparison of Dynamic vs Static Load Balancing Policies in a Mainframe – Personal Computer Network Model. *Technical Report CDC00-INV 1601*, University of Tsukuba.

Kameda, H., Altman, E. and Li, J. (2000). Paradoxes in Performance Optimization of Distributed Systems. *Proceedings of SSGRR 2000*. 18-28.

Karatza, H. D. and Hilzer, R. C. (2003). Performance Analysis of Parallel job Scheduling in Distributed systems. *Proceeding of the 36th annual Symposium on Simulation*.

Kleinrock, L. (1976). *Queueing Systems*. Vol.2, Computer Applications. Wiley, New York.

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.

Komaralingam, A. and Elleithy, K. M. (2002). Using a Queueing Model to Analyze the Performance of Web Servers. Citeseer.IST

Koole, G. (1999). On the Static Assignment to parallel Servers. *IEEE Transaction on Automatic Control* . 44, 588-1592.

Kouvatsos, D. D. (1986). A Maximum Entrophy Queue Length Distribution for A G/G/1 Finite Capacity Queue. *Journal of ACM*, 224-236.

Kouvatsos, D. D. and Othman, A. T. (1989a). Optimal Flow Control of a G/G/1 Queue. *International J. of Systems Science*. 20(2), 251-265.

Kouvatsos, D. D. and Othman, A. T. (1989b). Optimal Flow Control of a G/G/C Finite Capacity Queue. J. Operational Research Society. 40(7), 659-670.

Kouvatsos, D. D. and Othman, A. T. (1989c). Optimal Flow Control of end to end Packet Switched Network with Random Routing. *IEEE Proceeding*. 136(2), 90-100.

Krueger, P. and Shivaratri, N. G. (1994). Adaptive Location Policies for Global Scheduling, *IEEE Transactions on Software Engineering*, 20(6): 432 – 444.

Ku-Mahamud, K. R. (1993). Analysis and Decentralized Optimal Flow Control of Heterogeneous Computer Communication Ntework Models. *PhD thesis*. Universiti Pertanian Malaysia.

Kunz, T. (1991), The Influence of Different Workload Descriptions on a Heuristic Load Balancing Scheme. *IEEE Transactions on Software Engineering*, 17(7): 725-730.

Lawson, B. G. and Smirni, E. (2002). Multiple-queue Backfilling Scheduling with Priorities and Reservations for Parallel Systems. *ACM SIGMETRICS Performance Evaluation Review*, 29(4), 40-47.

Lazar, A. A. (1981). Optimal Control of an M/M/1 Queue. *In: Proc. 19th Allerton Conf. On Communication, Control and Computing*. 279-289.

Lazar, A. A. (1982). Centralized Optimal Control of a Jacksonian Network. *In Proceedings of the Sixteenth Conference on Information Science and Systems*. 17-19 March 1982, New Jersey, 316.

Lazar, A. A. (1983). The Throughput Time Delay Function of an M/M/1 Queue. *IEEE Transaction on Information Theory*, 6, 1001-1007.

Lazar, A. A. (1984). Optimal Control of an M/M/m Queue. *Journal of the Association for Computing Machinery*. 1(31), 86-98.

Lazowska, E. D., Zahorjan J., Graham, G. S. and Sevcik, K. C. (1984). *Quantitative System Performance. Computer System Analysis Using Queueing Network Models.* Prentice-Hall, New Jersey.

Leslie, R. and McKenzie, S. (1999). Evaluation of Loadsharing Algorithms for Heterogeneous Distributed Systems. *Computer Communication*, 22(4): 376 – 389.

Lin, W. and Kumar, A. (1984). Optimal Control of a Queueing System with Two Heterogeneous Servers. *IEEE Trans Automatic Control*, AC-29(8), 696-703.

Little, J. D. C. (1961). A proof of the Queueing Formula L = λW. *Operational Research 9*, 383-387.

Liu, J. B. (1999). A Multilevel Load Balancing Algorithm in a Distributed System. *Proceedings of the 19th annual conference on Computer Science*, 135-142.

Liu, C. (2000). Buffer Requirements for Zero Loss Flow Control with Explicit Congestion Notification, *Computer Communication, ICC 2000.*

Liu, C. and Jain, R. (2001). Improving Explicit Congestion Notification with Matk-Frong Strategy. *Computer Networks*, 35(3), 185-201.

Lublin, U. and Feitelson D. (2003). The workload on parallel supercomputers: modeling the characteristics of rigid jobs. *Journal of Parallel and Distributed Computing*, 63(11), 1105 - 1122 .

Menasce, D. A. and Almeida, V. A. F. (2000). *Scaling for E-Business.* New Jersey: Prentice Hall.

Nelson, R., and Towsley, D. (1985a). Comparison of Threshold Scheduling Policies for Multiple server systems. *IBM Research Report RC 11256.*

Nelson, R. and Towsley, D (1985b). *On Maximizing the Number of Departures Before a Deadline on Multiple Processors.* Technical Report RC 11256, IBM Thomas J. Watson Research Center, Yorktown Heights. New York.

Ni, L. M. and Hwang, K. (1981). Probabilistic Load Balancing in a Multiple Processor System with Many Job Classes. *Technical report RC 11256*, IBM Thomas J. Watson Research Center, Yorktown Height, New York..

Ni, L. M., and Hwang, K. (1985). Optimal Load Balancing in a Multiple Processor System with Many Job Classes. *IEEE Trans. Software Engineering*. 491-496.

Oudshoorn, M. J., and Huang, L. (1997). Conditional Task Scheduling on Loosely-Coupled Distributed Processors. *In The 10th International Conference on Parallel and Distributed Computer Systems,* 136-140.

Rahim, R. Othman, T., Isa, I. (2001). The Need For Business Process Modeling For Knowledge Representation, Knowledge Management 2001, *In Proceeding of International Conference and Exhibition*, Langkawi.

Rahim, R., Ku Mahamud, K. R., (2000). A Queueing Theory Approach On Workflow Performance, *In Proceeding of Allied Academies conference*, Maui, Hawaii.

Rahim, R., Othman, T., Ku Mahamud, K. R. (2001). Modeling and Performance Analysis of E-Procurement Workflow Using Petri Net, Malaysian Science and Technology Congress 2001 (MSTC 2001), *Symposium Information and Communication Technology*, Pulau Pinang.

Rahim, R., Othman, T., Ku-Mahamud, K. R. (2001)., Performance Analysis of E-Procurement Using Stochastic Petri Net, *Journal of Institute Mathematics & Computer Sciences (Computer Science Series)*, 12(2): 185-191.

Ramakrishnan, K. K. (1983). The Design and Analysis of Resource Allocation Policies in Distributed Systems. *Ph.D. Thesis*, University of Maryland, Dept. of Computer Science.

Reeser, P. and Hariharan, R. (2000). Analytic Model of Web Servers in Distributed Environments. *Proceedings of the second International workshop on Software and Performance..*32-41.

Ross, K. W. and Yao, D. D. (1991). Optimal Load Balancing and Scheduling. *Distributed Computer System Journal of the ACM (JACM)* 38 (3), 679-690.

Sasaki, G. (1989). Input Buffer Requirement for Round Robin Polling Systems. *In Proc. Allerton Conference on Communication, Control and Computing.*

Schopf, J. M. and Berman, F. (1999). Stochastic Scheduling. *In: Proceedings of Supercomputing '99.*

Schroeder, B. and Harchol-Balter, M. (2000). Evaluation of Task Assignment Policies for Supercomputing Servers: The case for load unbalancing and fairness. *Proceedings of the 9th. IEEE International symposium on High Performance Distributed Computing (HPDC).*

Sethuraman, J. and Squillante M. S. (1999). Optimal Stochastic Scheduling in Multiclass Parallel Queues. *ACM SIGMETRICS '99.* Atlanta, 93-125.

Shivaratri, N. G., Krueger, P. and Singhal, M. (1992). Load Distribution for Locally Distributed Systems. *IEEE Computer*, 8(12): 33-44.

Sih, G. C. and Lee, E. A. (1990a). Dynamic-level Scheduling for Heterogeneous Processor Networks. *In: Proceedings of the Second IEEE Symposium on Parallel and Distributed Systems*, 23-29.

Sih, G. C. and Lee, E. A.(1990b). Scheduling to Account for Interprocessor Communication within Interconnection-constrained Processor Networks. *In Proceedings of the 1990 International Conference on Parallel Processing*, 43-56.

Sih, G. C. and Lee, E. A (1993). A Compile-time Scheduling Heuristic for Interconnection-constrained Heterogeneous Processor Architectures. *IEEE Transactions on Parallel and Distributed Systems*, 4(2), 145-157.

Shenker, S. and Weinrib, A. (1989). The Optimal Control of Heterogeneous Queueing Systems: A paradigm for load sharing and routing. *IEEE Transactions on Computers*, 38(12), 1724-1735.

Shioyama, T. (1991). Optimal Control of Queueing System with Two Types of Customers. *European Journal of Operational Research.* 52, 367-372.

Smith, C. U. and Williams, L. G. (2001). *Performance Solutions, A Practical Guide to Creating Responsive, Scalable Software.* Indianapolis: Pearson Education.

Stidham, S. JR. (1985). Optimal Control of Admission to a Queueing system. *IEEE Trans on Automatic Control,* AC-30(8), 705-713.

Tang, X. and Chanson, S. T., (2000). Optimizing Static Job Scheduling in a Network of Heterogeneous Computers. *In Proceedings of the 29th. International Conference on Parallel Processing (ICPP),* 373 – 382.

Tantawi, A. N., and Towsley, D. (1985). Optimal Static Load Balancing in Distributed Computer Systems, *J. ACM.* 32(2), 445-465.

Tavana, M. and Rappaport, J. (1997). Optimal Allocation of Arrivals to a Collection of Parallel Workstations. *International Journal of Operations & Production Management,* 17(3), 305-325.

Waldspurger, C. A. and Weihl. W. E. (1994). Lottery scheduling: Flexible Proportional-Share Resource Management. *In Proceedings of the First Symposium on Operating System Design and Implementation.* 212-225.

Wang, Y-T., Morris, R. J. T. (1985). Load Sharing in Distributed Sistems. *IEEE Trans. Computers* C-34, 204-217.

Weinrib, A. and Shenker, S. (1988). Greed is not enough: Adaptive Load Sharing in Large Heterogeneous Systems. *IEEE INFOCOM'88.* The Conference on Computer Communications Proceedings, 986-994.

Winston, W. (1977). Optimality of The Shortest Line Discipline. *SIAM J. Appl. Prob.* 14, 181-189.

Wolf, J. L., Turek, J., Chen, M., and Yu, P. S. (1994). *ACM SIGMETRICS,* 34-46.

Wolf, J. L. and Yu, P. S. (2001). On Balancing the Load in a Clustered Web Farm. *ACM Transactions on Internet Technology,* 1(2), 231-261.

Yao, D. D. (1994). *Stochastic Modeling and Analysis of Manufacturing Systems.* Springer-Verlag, Berlin.

Zaki, M. J. Li, W. and Srinivasan, P. (1996). Customized Dynamic Load Balancing for Network of Work-stations. *In Proceedings of HPDC '96.* 17-28.

Zahorjan, A. J. (1983). Workload representations in queueing models of computer systems. Proceedings of the 1983 *ACM SIGMETRICS* Conference on Measurement and Modeling of Computer systems, 70-81

Zahorjan, A. J. (1992). Scheduling a Mixed Interactive and Batch Workload on a Parallel, Shared Memory Supercomputer. *Supercomputing 92.*