MATRIX ANALYTICAL METHODS FOR RETRIAL QUEUES

by

Jeffrey E. Diamond

A Thesis

Submitted to the Faculty of Graduate Studies

in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

Department of Mechanical and Industrial Engineering

University of Manitoba

Winnipeg, Manitoba

(c) September, 1995

National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

The author has granted an irrevocable non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of his/her thesis by any means and in any form or format, making this thesis available to interested persons.

L'auteur a accordé une licence irrévocable et non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de sa thèse de quelque manière et sous quelque forme que ce soit pour mettre des exemplaires de cette thèse à la disposition des personnes intéressées.

The author retains ownership of the copyright in his/her thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without his/her permission.

L'auteur conserve la propriété du droit d'auteur qui protège sa thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

ISBN 0-612-13077-0

Canada

Name _____

Dissertation Abstracts International and *Masters Abstracts International* are arranged by broad, general subject categories. Please select the one subject which most nearly describes the content of your dissertation or thesis. Enter the corresponding four-digit code in the spaces provided.

Operations Research

SUBJECT TERM

0796

SUBJECT CODE

UMI

Subject Categories

THE HUMANITIES AND SOCIAL SCIENCES

COMMUNICATIONS AND THE ARTS

Architecture 0729
Art History 0377
Cinema 0900
Dance 0378
Fine Arts 0357
Information Science 0723
Journalism 0391
Library Science 0399
Mass Communications 0708
Music 0413
Speech Communication 0459
Theater 0465

EDUCATION

General 0515
Administration 0514
Adult and Continuing 0516
Agricultural 0517
Art 0273
Bilingual and Multicultural 0282
Business 0688
Community College 0275
Curriculum and Instruction 0727
Early Childhood 0518
Elementary 0524
Finance 0277
Guidance and Counseling 0519
Health 0680
Higher 0745
History of 0520
Home Economics 0278
Industrial 0521
Language and Literature 0279
Mathematics 0280
Music 0522
Philosophy of 0998
Physical 0523
Psychology 0525
Reading 0535
Religious 0527
Sciences 0714
Secondary 0533
Social Sciences 0534
Sociology of 0340
Special 0529
Teacher Training 0530
Technology 0710
Tests and Measurements 0288
Vocational 0747

LANGUAGE, LITERATURE AND LINGUISTICS

Language
- General 0679
- Ancient 0289
- Linguistics 0290
- Modern 0291

Literature
- General 0401
- Classical 0294
- Comparative 0295
- Medieval 0297
- Modern 0298
- African 0316
- American 0591
- Asian 0305
- Canadian (English) 0352
- Canadian (French) 0355
- English 0593
- Germanic 0311
- Latin American 0312
- Middle Eastern 0315
- Romance 0313
- Slavic and East European 0314

PHILOSOPHY, RELIGION AND THEOLOGY

Philosophy 0422
Religion
- General 0318
- Biblical Studies 0321
- Clergy 0319
- History of 0320
- Philosophy of 0322

Theology 0469

SOCIAL SCIENCES

American Studies 0323
Anthropology
- Archaeology 0324
- Cultural 0326
- Physical 0327

Business Administration
- General 0310
- Accounting 0272
- Banking 0770
- Management 0454
- Marketing 0338

Canadian Studies 0385
Economics
- General 0501
- Agricultural 0503
- Commerce-Business 0505
- Finance 0508
- History 0509
- Labor 0510
- Theory 0511

Folklore 0358
Geography 0366
Gerontology 0351
History
- General 0578
- Ancient 0579
- Medieval 0581
- Modern 0582
- Black 0328
- African 0331
- Asia, Australia and Oceania 0332
- Canadian 0334
- European 0335
- Latin American 0336
- Middle Eastern 0333
- United States 0337

History of Science 0585
Law 0398
Political Science
- General 0615
- International Law and Relations 0616
- Public Administration 0617

Recreation 0814
Social Work 0452
Sociology
- General 0626
- Criminology and Penology ... 0627
- Demography 0938
- Ethnic and Racial Studies 0631
- Individual and Family Studies 0628
- Industrial and Labor Relations 0629
- Public and Social Welfare 0630
- Social Structure and Development 0700
- Theory and Methods 0344

Transportation 0709
Urban and Regional Planning 0999
Women's Studies 0453

THE SCIENCES AND ENGINEERING

BIOLOGICAL SCIENCES

Agriculture
- General 0473
- Agronomy 0285
- Animal Culture and Nutrition 0475
- Animal Pathology 0476
- Food Science and Technology 0359
- Forestry and Wildlife 0478
- Plant Culture 0479
- Plant Pathology 0480
- Plant Physiology 0817
- Range Management 0777
- Wood Technology 0746

Biology
- General 0306
- Anatomy 0287
- Biostatistics 0308
- Botany 0309
- Cell 0379
- Ecology 0329
- Entomology 0353
- Genetics 0369
- Limnology 0793
- Microbiology 0410
- Molecular 0307
- Neuroscience 0317
- Oceanography 0416
- Physiology 0433
- Radiation 0821
- Veterinary Science 0778
- Zoology 0472

Biophysics
- General 0786
- Medical 0760

EARTH SCIENCES

Biogeochemistry 0425
Geochemistry 0996
Geodesy 0370
Geology 0372
Geophysics 0373
Hydrology 0388
Mineralogy 0411
Paleobotany 0345
Paleoecology 0426
Paleontology 0418
Paleozoology 0985
Palynology 0427
Physical Geography 0368
Physical Oceanography 0415

HEALTH AND ENVIRONMENTAL SCIENCES

Environmental Sciences 0768
Health Sciences
- General 0566
- Audiology 0300
- Chemotherapy 0992
- Dentistry 0567
- Education 0350
- Hospital Management 0769
- Human Development 0758
- Immunology 0982
- Medicine and Surgery 0564
- Mental Health 0347
- Nursing 0569
- Nutrition 0570
- Obstetrics and Gynecology .. 0380
- Occupational Health and Therapy 0354
- Ophthalmology 0381
- Pathology 0571
- Pharmacology 0419
- Pharmacy 0572
- Physical Therapy 0382
- Public Health 0573
- Radiology 0574
- Recreation 0575
- Speech Pathology 0460
- Toxicology 0383

Home Economics 0386

PHYSICAL SCIENCES

Pure Sciences

Chemistry
- General 0485
- Agricultural 0749
- Analytical 0486
- Biochemistry 0487
- Inorganic 0488
- Nuclear 0738
- Organic 0490
- Pharmaceutical 0491
- Physical 0494
- Polymer 0495
- Radiation 0754

Mathematics 0405
Physics
- General 0605
- Acoustics 0986
- Astronomy and Astrophysics 0606
- Atmospheric Science 0608
- Atomic 0748
- Electronics and Electricity 0607
- Elementary Particles and High Energy 0798
- Fluid and Plasma 0759
- Molecular 0609
- Nuclear 0610
- Optics 0752
- Radiation 0756
- Solid State 0611

Statistics 0463

Applied Sciences

Applied Mechanics 0346
Computer Science 0984
Engineering
- General 0537
- Aerospace 0538
- Agricultural 0539
- Automotive 0540
- Biomedical 0541
- Chemical 0542
- Civil 0543
- Electronics and Electrical 0544
- Heat and Thermodynamics ... 0348
- Hydraulic 0545
- Industrial 0546
- Marine 0547
- Materials Science 0794
- Mechanical 0548
- Metallurgy 0743
- Mining 0551
- Nuclear 0552
- Packaging 0549
- Petroleum 0765
- Sanitary and Municipal 0554
- System Science 0790

Geotechnology 0428
Operations Research 0796
Plastics Technology 0795
Textile Technology 0994

PSYCHOLOGY

General 0621
Behavioral 0384
Clinical 0622
Developmental 0620
Experimental 0623
Industrial 0624
Personality 0625
Physiological 0989
Psychobiology 0349
Psychometrics 0632
Social 0451

Nom _____

Dissertation Abstracts International est organisé en catégories de sujets. Veuillez s.v.p. choisir le sujet qui décrit le mieux votre thèse et inscrivez le code numérique approprié dans l'espace réservé ci-dessous.

_____ ☐☐☐ U·M·I

SUJET — CODE DE SUJET

Catégories par sujets

HUMANITÉS ET SCIENCES SOCIALES

COMMUNICATIONS ET LES ARTS
Architecture 0729
Beaux-arts 0357
Bibliothéconomie 0399
Cinéma 0900
Communication verbale 0459
Communications 0708
Danse 0378
Histoire de l'art 0377
Journalisme 0391
Musique 0413
Sciences de l'information 0723
Théâtre 0465

ÉDUCATION
Généralités 515
Administration 0514
Art 0273
Collèges communautaires 0275
Commerce 0688
Économie domestique 0278
Éducation permanente 0516
Éducation préscolaire 0518
Éducation sanitaire 0680
Enseignement agricole 0517
Enseignement bilingue et multiculturel 0282
Enseignement industriel 0521
Enseignement primaire. 0524
Enseignement professionnel 0747
Enseignement religieux 0527
Enseignement secondaire 0533
Enseignement spécial 0529
Enseignement supérieur 0745
Évaluation 0288
Finances 0277
Formation des enseignants 0530
Histoire de l'éducation 0520
Langues et littérature 0279
Lecture 0535
Mathématiques 0280
Musique 0522
Orientation et consultation 0519
Philosophie de l'éducation 0998
Physique 0523
Programmes d'études et enseignement 0727
Psychologie 0525
Sciences 0714
Sciences sociales 0534
Sociologie de l'éducation 0340
Technologie 0710

LANGUE, LITTÉRATURE ET LINGUISTIQUE
Langues
- Généralités 0679
- Anciennes 0289
- Linguistique 0290
- Modernes 0291

Littérature
- Généralités 0401
- Anciennes 0294
- Comparée 0295
- Médiévale 0297
- Moderne 0298
- Africaine 0316
- Américaine 0591
- Anglaise 0593
- Asiatique 0305
- Canadienne (Anglaise) 0352
- Canadienne (Française) 0355
- Germanique 0311
- Latino-américaine 0312
- Moyen-orientale 0315
- Romane 0313
- Slave et est-européenne 0314

PHILOSOPHIE, RELIGION ET THÉOLOGIE
Philosophie 0422
Religion
- Généralités 0318
- Clergé 0319
- Études bibliques 0321
- Histoire des religions 0320
- Philosophie de la religion 0322

Théologie 0469

SCIENCES SOCIALES
Anthropologie
- Archéologie 0324
- Culturelle 0326
- Physique 0327

Droit 0398
Économie
- Généralités 0501
- Commerce-Affaires 0505
- Économie agricole 0503
- Économie du travail 0510
- Finances 0508
- Histoire 0509
- Théorie 0511

Études américaines 0323
Études canadiennes 0385
Études féministes 0453
Folklore 0358
Géographie 0366
Gérontologie 0351
Gestion des affaires
- Généralités 0310
- Administration 0454
- Banques 0770
- Comptabilité 0272
- Marketing 0338

Histoire
- Histoire générale 0578
- Ancienne 0579
- Médiévale 0581
- Moderne 0582
- Histoire des noirs 0328
- Africaine 0331
- Canadienne 0334
- États-Unis 0337
- Européenne 0335
- Moyen-orientale 0333
- Latino-américaine 0336
- Asie, Australie et Océanie 0332

Histoire des sciences 0585
Loisirs 0814
Planification urbaine et régionale 0999
Science politique
- Généralités 0615
- Administration publique 0617
- Droit et relations internationales 0616

Sociologie
- Généralités 0626
- Aide et bien-être social 0630
- Criminologie et établissements pénitentiaires 0627
- Démographie 0938
- Études de l' individu et de la famille 0628
- Études des relations interethniques et des relations raciales 0631
- Structure et développement social 0700
- Théorie et méthodes. 0344
- Travail et relations industrielles 0629

Transports 0709
Travail social 0452

SCIENCES ET INGÉNIERIE

SCIENCES BIOLOGIQUES
Agriculture
- Généralités 0473
- Agronomie. 0285
- Alimentation et technologie alimentaire 0359
- Culture 0479
- Élevage et alimentation 0475
- Exploitation des pâturages 0777
- Pathologie animale 0476
- Pathologie végétale 0480
- Physiologie végétale 0817
- Sylviculture et faune 0478
- Technologie du bois 0746

Biologie
- Généralités 0306
- Anatomie 0287
- Biologie (Statistiques) 0308
- Biologie moléculaire 0307
- Botanique 0309
- Cellule 0379
- Écologie 0329
- Entomologie 0353
- Génétique 0369
- Limnologie 0793
- Microbiologie 0410
- Neurologie 0317
- Océanographie 0416
- Physiologie 0433
- Radiation 0821
- Science vétérinaire 0778
- Zoologie 0472

Biophysique
- Généralités 0786
- Medicale 0760

SCIENCES DE LA TERRE
Biogéochimie 0425
Géochimie 0996
Géodésie 0370
Géographie physique 0368
Géologie 0372
Géophysique 0373
Hydrologie 0388
Minéralogie 0411
Océanographie physique 0415
Paléobotanique 0345
Paléoécologie 0426
Paléontologie 0418
Paléozoologie 0985
Palynologie 0427

SCIENCES DE LA SANTÉ ET DE L'ENVIRONNEMENT
Économie domestique 0386
Sciences de l'environnement 0768
Sciences de la santé
- Généralités 0566
- Administration des hôpitaux 0769
- Alimentation et nutrition 0570
- Audiologie 0300
- Chimiothérapie 0992
- Dentisterie 0567
- Développement humain 0758
- Enseignement 0350
- Immunologie 0982
- Loisirs 0575
- Médecine du travail et thérapie 0354
- Médecine et chirurgie 0564
- Obstétrique et gynécologie 0380
- Ophtalmologie 0381
- Orthophonie 0460
- Pathologie 0571
- Pharmacie 0572
- Pharmacologie 0419
- Physiothérapie 0382
- Radiologie 0574
- Santé mentale 0347
- Santé publique 0573
- Soins infirmiers 0569
- Toxicologie 0383

SCIENCES PHYSIQUES
Sciences Pures
Chimie
- Généralités 0485
- Biochimie 487
- Chimie agricole 0749
- Chimie analytique 0486
- Chimie minérale 0488
- Chimie nucléaire 0738
- Chimie organique 0490
- Chimie pharmaceutique 0491
- Physique 0494
- PolymÈres 0495
- Radiation 0754

Mathématiques 0405
Physique
- Généralités 0605
- Acoustique 0986
- Astronomie et astrophysique 0606
- Électronique et électricité 0607
- Fluides et plasma 0759
- Météorologie 0608
- Optique 0752
- Particules (Physique nucléaire) 0798
- Physique atomique 0748
- Physique de l'état solide 0611
- Physique moléculaire 0609
- Physique nucléaire 0610
- Radiation 0756

Statistiques 0463

Sciences Appliqués Et Technologie
Informatique 0984
Ingénierie
- Généralités 0537
- Agricole 0539
- Automobile 0540
- Biomédicale 0541
- Chaleur et thermodynamique 0348
- Conditionnement (Emballage) 0549
- Génie aérospatial 0538
- Génie chimique 0542
- Génie civil 0543
- Génie électronique et électrique 0544
- Génie industriel 0546
- Génie mécanique 0548
- Génie nucléaire 0552
- Ingénierie des systèmes 0790
- Mécanique navale 0547
- Métallurgie 0743
- Science des matériaux 0794
- Technique du pétrole 0765
- Technique minière 0551
- Techniques sanitaires et municipales 0554
- Technologie hydraulique 0545

Mécanique appliquée 0346
Géotechnologie 0428
Matières plastiques (Technologie) 0795
Recherche opérationnelle 0796
Textiles et tissus (Technologie) 0794

PSYCHOLOGIE
Généralités 0621
Personnalité 0625
Psychobiologie 0349
Psychologie clinique 0622
Psychologie du comportement 0384
Psychologie du développement 0620
Psychologie expérimentale 0623
Psychologie industrielle 0624
Psychologie physiologique 0989
Psychologie sociale 0451
Psychométrie 0632

MATRIX ANALYTICAL METHODS FOR RETRIAL QUEUES

BY

JEFFREY E. DIAMOND

A Thesis submitted to the Faculty of Graduate Studies of the University of Manitoba in partial fulfillment of the requirements of the degree of

DOCTOR OF PHILOSOPHY

© 1996

Permission has been granted to the LIBRARY OF THE UNIVERSITY OF MANITOBA to lend or sell copies of this thesis, to the NATIONAL LIBRARY OF CANADA to microfilm this thesis and to lend or sell copies of the film, and LIBRARY MICROFILMS to publish an abstract of this thesis.

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

ABSTRACT

Matrix analytical methods are developed for the modelling and analysis of some retrial queues. A sufficient condition for ergodicity of a single server retrial queue with Markovian arrival process and phase type service is derived and block Gaussian elimination is used to solve linear systems of equations arising in the calculation of the stationary distribution of states, the distribution of the number of retrials executed by an arbitrary customer, the waiting time distribution and the moments of the waiting time. A bound is obtained on the probability lost due to truncation of the infinite generator by considering an approximation which stochastically dominates the exact queue. For the special case of Poisson arrivals, an explicit expression is obtained for the level dependent rate matrices and a sufficient condition for ergodicity is obtained by considering the convergence of a related matrix series. Efficient numerical methods are developed for retrial queues with finite buffers or multiple servers and an approximate model is developed for retrial queues with phase type interretrial times. A sufficient condition for ergodicity is obtained and numerical experiments are performed to examine the effectiveness of the approximation in predicting the first two moments of the waiting time. A level dependent extension of the single server retrial queue with Markovian arrival process and phase type service is used to model a local area network with CSMA protocol and extensions of the model with Poisson arrivals, arbitrary level dependence and geometric loss are considered.

ACKNOWLEDGEMENTS

This research has been supported in part by a Duff Roblin Post-Graduate Scholarship from the University of Manitoba.

I thank my advisor, Dr. Alfa, for introducing me to the subject of queueing theory with enthusiasm.

I give credit to my family for the support that I have recieved over the many years of my study: To my children for their patience and to my wife for her love and understanding. I would also like to thank my parents and my grandparents for instilling in me a healthy respect for education and for their conviction that I could make it this far.

TABLE OF CONTENTS

INTRODUCTION

A retrial queue is one in which a customer, if he finds all servers busy and waiting positions (if any) occupied upon arrival, waits some random period of time (usually exponentially distributed) and then tries for service again. While the customer is waiting to retry he is considered to be "in orbit "and retries for service periodically at random intervals until he either seizes a server or waiting position or gives up and is lost to the system forever. These models are common in areas such as telephone and computer communication systems. There is a large volume of research which has been published on various models of retrial queues and comprehensive reviews can be found in Yang and Templeton (1987) and Falin (1990).

Results are available for retrial queues with loss (customers may give up and leave the system), with batch arrivals, with finite buffers, and with multiple servers as well other structurally complex systems. Keilson et al. (1968) considered the M/G/1 retrial queue and used the method of supplementary variables to obtain the generating function for the joint distribution of the number of customers in the system and elapsed service time. Falin (1979) obtained the joint Laplace transform of the length of the busy period and the number of customers served in the period. He also obtained (1991) the Laplace transform of the waiting time for the M/G/1 queue. Yang (1990) developed algoithms to evaluate the stationary distribution of queue length for the GI/M/s/m retrial queue.

So far, most models of retrial queues have been analyzed using classical analytic methods which are sometimes difficult and risky in their numerical implementation. Results are usually given in the form of Laplace transforms or generating functions which must be obtained by solving an integral and then inverted . Little or no attention is paid

to the algorithmic implementation of these solutions. While the direct numerical computation of existing analytic solutions is best considered within the framework of classical numerical analysis, there is clearly room for an approach in which algorithmic feasibility is a primary criterion in the modelling process from the outset. The field of computational probability and, more specifically, matrix analytical methods adopt this point of view. Matrix analytical methods take advantage of structural properties of Markov chains in order to obtain stable numerical methods for the calculation of quantities of interest. They also provide insight into the behavior of models through a probabilistic understanding of intermediate steps of algorithms which is unavailable in the purely formal manipulations of many classical methods.

Until now, all models of retrial queues have assumed either exponentially distributed service times or a Poisson input process. While these assumptions have been necessary in order to make the analysis of the models tractable, they may not accurately reflect the behavior of real systems of interest. The matrix analytical approach allows us to avoid these restrictive assumptions. If we replace the Poisson process and the exponential distribution with the more general Markovian arrival process (MAP) and phase type (PH) distribution (see Neuts 1981 and Lucantoni 1991), we can obtain queues, with neither Poisson input nor exponential service times, which can be modelled as Markov Chains. We can then exploit the structure of the Markov chains to obtain stable and efficient numerical methods for the analysis of the queueing models.

A phase type distribution is the distribution of the time until absorption in a finite absorbing Markov chain and is represented by a pair of objects of the form (β^T, S). β^T is the initial probability vector for the absorbing chain and S is the portion of the generator for the absorbing chain corresponding to transitions between transient states. Any distribution on $(0, \infty)$ can, in principle, be approximated arbitrarily closely

by a phase type distribution, although the dimensions of β^T and S may be large. Phase type distributions of reasonably small dimensions have been used to approximate many distributions of practical importance (see Bobbio et al. 1980, Johnson and Taaffe 1988, Asmussen and Nerman 1991) and some very common distributions such as the Erlang, Coxian and hyperexponential are special cases of the phase type distribution.

The Markovian arrival process generalizes the phase type renewal process by allowing the initial probability vector for each phase type interarrival time to depend on the phase (transient state of absorbing chain) from which the last interarrival time terminated (entered absorbing state). This allows for correlation between successive interarrival times and we can use MAPs to model processes which are bursty.

In the first chapter, we consider the MAP/PH/1 retrial queue. We apply block gaussian elimination to obtain the stationary distribution of the number of customers in the orbit and the number of retrials performed by an arbitrary customer as well as the moments of the waiting time. We furnish probabilistic interpretation for all of the objects involved in the calculation and we obtain a bound for the probability lost to truncation of the infinite dimensional generator. We apply the randomization method to obtain the distribution of the waiting time and suggest an approximation for the waiting time distribution, based on the distribution of the number of retrials, which is less computationally intensive. We also present a state dependent extension to the MAP/PH/1 retrial queue which models a local area network with carrier sense multiple access protocol.

In the second chapter, we obtain some explicit results for the special case of Poisson arrivals and a sufficient (stability) condition for convergence of a matrix series which arises in the normalization of the stationary probability vector. We also consider the possibility of state dependent extensions and examine another extension, the M/M/1 re-

trial queue with los, from the matrix analytical point of view. In Chapter Three, we adapt the methods of the first chapter to retrial queues with finite buffers or multiple servers. In Chapter Four, we adapt an approximation method developed by Yang et al. (1994) for the stationary distribution of M/G/1 retrial queues with general retrial times to the problem of determining the waiting time distribution and moments and the distribution of the number of retrials for M/PH/1 queues with phase type retrial times. We perform some numerical experiments to determine the accuracy of the approximation in determining the first two moments of the waiting time.

CHAPTER ONE

THE MAP/PH/1 RETRIAL QUEUE

1.1 INTRODUCTION

So far, most models of retrial queues have employed the Poisson arrival proces to model the input stream to the queue. This is somewhat restrictive since arrival streams to various kinds of queues are often not well modelled by a Poisson process. Homitchkov (1987 and 1988) considered retrial queues with a single exponential server and independent interarrival times with Erlangian or hyperexponential distributions of second order. Yang (1990) obtained the stationary distribution of queue size for the GI/M/s/m retrial queue and Yang, Posner and Templeton (1992) developed numerical methods for the special case of Coxian arrival process.

Sondermann and Pourbabai (1987) developed an approximation method for single server recirculation systems. These systems are similar to retrial queues in that there are no waiting spaces and the overflow of customers not receiving service merges with the arrival process from outside to request service again instead of disappearing from the system. The algorithm used is an iterative one in which, at each stage, the overflow process is approximated by matching its first two moments and then merged with the arrival process from outside. The resulting superposition arrival process is then approximated by matching its first two moments and becomes the new input stream (to the server) used to recalculate the overflow moments. The interarrival and service time distributions are assumed to be either hyperexponential of order two or delayed exponential for the purpose of moment matching but the algorithm is intended to be applied to the case of arbitrary distributions for

these times.

We consider an exact model of a retrial queue with non-exponential service times as well as non-exponential interarrival times. Service and interarrival time distributions are assumed to be members of the class PH of phase type distributions defined by Neuts (1981). This is a versatile class of distributions and can be used to approximate any distribution arbitrarily closely. The phase type distribution with representation (β, S) is the distribution of the time until absorption of the finite dimensional absorbing Markov chain with generator

$$S' = \begin{bmatrix} S^0 & S \\ 0 & 0 \end{bmatrix}$$

and initial probability vector $(0, \beta^T)$ where $S^0 = -Se$ and e is a column vector of 1's. The subgenerator S has negative diagonal elements and nonnegative off-diagonal elements and S^0 is nonnegative. The cumulative distribution function of the time until absorption is given by $F(t) = 1 - \beta^T exp[St]e$.

Since a superposition of renewal processes is not , in general , a renewal process, considering only arrival streams which are renewal processes may be too restrictive. The arrival process may be bursty and successive interarrival times may be correlated. The Markovian arrival process (MAP) provides a model which posesses these characteristics and which is tractable via matrix analytical methods. It is a rich class of processes which includes, as special cases, the phase type renewal process and the Markov modulated Poisson process. A Markovian arrival process can be considered a Markov process $\{N(t), J(t)\}$ on the state space $\{(i,j) : i \geq 0, 1 \leq j \leq m\}$ with an infinitesimal generator Q having the structure

$$Q = \begin{bmatrix} D_0 & D_1 & 0 & \cdots & & \\ 0 & D_0 & D_1 & 0 & \cdots & \\ 0 & 0 & D_0 & D_1 & 0 & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \end{bmatrix}$$

where D_0 and D_1 are $m \times m$ matrices, D_0 has negative diagonal elements and nonnegative off-diagonal elements, D_1 is nonnegative and $(D_0 + D_1)e = 0$ where e is an m dimensional column vector of ones. $N(t) = i$ represents the number of arrivals in $(0, t)$ and $J(t) = j$ represents an auxiliary state or phase variable.

We consider a MAP/PH/1 retrial queue with arrival process represented by (D_0, D_1), service time distribution represented by (β^T, S) and interretrial times exponentially distributed with rate θ. The state space is given by $\{(i, j, k, \ell) | i = 0, 1, \ldots; j = 0, 1; k = 1, \ldots, m; \ell = 1, \ldots, n_j\}$ where $n_0 = 1$, $n_1 = n$ and m and n are the dimensions of the arrival and service process representations respectively. i, j, k, and ℓ represent the number of customers in orbit, state of server (idle for $k = 0$ and busy for $k = 1$), arrival phase and service phase respectively. If we order the state space so that the labels (i, j, k, ℓ) appear in lexicographic order, the generator has the following structure:

$$Q = \begin{bmatrix} A_{01} & A_{00} & 0 & 0 & \ldots \\ A_{12} & A_{11} & A_{10} & 0 & \ldots \\ 0 & A_{22} & A_{21} & A_{20} & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{1.1}$$

where:

$$A_{i0} = A_0 = \begin{bmatrix} 0 & 0 \\ 0 & D_1 \otimes I \end{bmatrix} \qquad A_{i1} = \begin{bmatrix} D_0 - i\theta I & D_1 \otimes \beta^T \\ I \otimes S^0 & I \otimes S + D_0 \otimes I \end{bmatrix}$$

$$A_{i2} = \begin{bmatrix} 0 & i\theta I \otimes \beta^T \\ 0 & 0 \end{bmatrix} \tag{1.2}.$$

and $S^0 = -Se$.

We are interested in the stationary distribution x which satisfies $xQ = 0$. We employ block Gaussian elimination to obtain a numerical solution. Because the subdiagonal blocks are quite sparse, we can cut down the complexity of each step of the algorithm to $O(m^3 n^2)$ from the

$O(m^3n^3)$ which would be required otherwise, where m and n are the dimensions of the arrival and service process representations respectively. We eliminate the levels (level i refers the set of states with i customers in the orbit.) $0, 1, 2, \ldots$ one at a time in that order. Eliminating the lower levels first allows us to defer deciding at what level to truncate or approximate the system until we have sufficient information to do so.

When applying numerical solutions to infinite systems with level dependence it is usually necessary to approximate the system either with a finite system or with a system which is spatially homogeneous for sufficiently high levels. We will employ the former method, however we will consider two spatially homogeneous extensions in order to determine where to truncate the system.

The first homogeneous extension we consider was proposed by Falin (1983) for the M/M/s retrial queue. In this approximation, the server begins a new service immediately after each service completion if there are more than a fixed number N of customers in the orbit. Thus, above level N , the queue is identical to a MAP/PH/1 queue with infinite buffer space and random service order and the stationary distribution has a modified matrix geometric (see Neuts 1981) form. The advantage of this approximation is that the rate matrix associated with the homogeneous extension is independent of the transition level N so that we can increase N without having to recalculate the rate matrix.

The second homogeneous extension we consider is similar to the one used by Neuts and Rao (1990) as an approximation to the M/M/s retrial queue. The approximation is obtained by allowing only N customers to attempt retrials when there are more than N customers in the orbit. The total retrial rate from the orbit is thus bounded by $N\theta$ and the resulting generator is spatially homogeneous above level N and has a modified matrix geometric stationary distribution. This approximation was also proposed by Greenberg (1986) and Stepanov (1988).

It is intuitively clear that this approximate system is , in some sense, less efficient than the exact queue since the mean server idle time is larger for each level. We can formalize this observation in terms of stochastic dominance of one process over another if one of the processes is monotonic (see Massey (1987)). We define a class of service time distributions which afford the queue the required monotonicity. Since the approximation is homogeneous above a certain level we can use the methods of Neuts (1981) to calculate the tail probability above a certain level. This provides an upper bound on the probability which is lost in the exact queue if we truncate at that level.

1.2 STABILITY CONDITION

We obtain a sufficient condition for ergodicity by applying Mustafa's criterion, as suggested in Falin (1984). A Markov chain $\{Z_n | n = 0, 1, \ldots\}$ on a state space Σ is ergodic if there exists a non-negative function $f : \Sigma \rightarrow \Re$ such that the mean drift $E[f(Z_{n+1})|Z_n = z] - f(z) < -\epsilon$ for some $\epsilon > 0$ and for all but a finite number of points $z \in \Sigma$

Proposition 1.1: *If a MAP/PH/1 queue with infinite buffer capacity and irreducible arrival and service process representations is ergodic, then the MAP/PH/1 retrial queue with identical arrival and service processes is also ergodic.*

Proof: . Let (D_0, D_1) and (β^T, S) be irreducible representations of the arrival and service processes respectively for an ergodic MAP/PH/1 queue. The transition probability matrix P for the jump chain of the

retrial queue imbedded at each event has the form

$$P = \begin{bmatrix} P_{01} & P_0 & 0 & 0 & \dots \\ P_{12} & P_{11} & P_0 & 0 & \dots \\ 0 & P_{22} & P_{21} & P_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

where

$$P_{i1} = \begin{bmatrix} I + (\Delta_0 + i\theta I)^{-1}(D_0 - i\theta I) & (\Delta_0 + i\theta I)^{-1}(D_1 \otimes \beta^T) \\ B_2 e & B_1 \end{bmatrix}$$

$$P_0 = \begin{bmatrix} 0 & 0 \\ 0 & B_0 \end{bmatrix} \qquad P_{i2} = \begin{bmatrix} 0 & i\theta(\Delta_0 + i\theta I)^{-1} \otimes \beta^T \\ 0 & 0 \end{bmatrix}$$

$$B_0 = \Delta_1^{-1}(D_1 \otimes I) \qquad B_1 = I + \Delta_1^{-1}(I \otimes S + D_0 \otimes I)$$

$$B_2 = \Delta_1^{-1}(I \otimes S^0 \beta^T)$$

$$\Delta_0 = -diag(D_0) \qquad \Delta_1 = -diag(I \otimes S + D_0 \otimes I).$$

Let R be the rate matrix associated with the infinite buffer queue and let $\eta < 1$ be its spectral radius. Then, from the discussion in the proof of Lemma 1.3.4 in Neuts (1981), we know that for all $\epsilon \in (\eta, 1)$, $sp(B_0 + \epsilon B_1 + \epsilon^2 B_2) = \epsilon' < \epsilon$. Let $x > 0$ be the right eigenvector associated with ϵ' so that $(B_0 + \epsilon B_1 + \epsilon^2 B_2)x = \epsilon' x < \epsilon x$.

Now define the vector $f = [f_0, f_1, f_2, \dots]$ according to

$$f_i = \epsilon^{-i} \left(\begin{bmatrix} \epsilon I \otimes \beta^T \\ I \end{bmatrix} x + ae \right).$$

where $a \in (0, 1)$. We need only show that $(P - I)f < -\epsilon'' e$ (except for a finite number of components) for some $\epsilon'' > 0$. Now we have

$$[(P - I)f]_i = \begin{bmatrix} \alpha_{i0} \\ \alpha_{i1} \end{bmatrix}$$

where

$$\alpha_{i0} = \epsilon^{-i}(\Delta_0 + i\theta I)^{-1} \left(\epsilon(D_0 + D_1)(I \otimes \beta^T)x - ia(1 - \epsilon)\theta e \right)$$

$$\alpha_{i1} = \epsilon^{-(i+1)} \left(a(1-\epsilon)B_0 e + (B_0 + \epsilon B_1 + \epsilon^2 B_2 - \epsilon I)x \right).$$

Since $\beta_1 = -(B_0 + \epsilon B_1 + \epsilon^2 B_2 - \epsilon I)x > 0$, we can choose $a \in (0,1)$ so that $\alpha_{i1} < -\epsilon'' e = -min\{\beta_1\}e/2$ for all $i \geq 1$. Since $\alpha_{i0} \to -\infty$ as $i \to \infty$ there exists an integer i' such that $\alpha_{i0} < -\epsilon'' e$ for al $i > i'$. Thus $[(P-I)f]_i < -\epsilon'' e$ for all $i > i'$. □

A necessary and sufficient condition for ergodicity of the MAP/PH/1 queue with infinite buffer capacity can be obtained by applying Theorem 1.7.1 in Neuts (1981). The condition is given by $\pi(D_1 e \otimes e - e \otimes S^0) < 0$ where π is the solution to $\pi[I \otimes (S + S^0\beta^T) + (D_0 + D_1) \otimes I] = 0$ and $\pi e = 1$.

We would also like to note, as a corollary to Proposition 1.1, that if a MAP/PH/1 queue with infinite buffer capacity is irreducible and ergodic, then the approximation to the retrial queue referred to as the second homogeneous extension (with bounded retrial rate) is also ergodic provided the maximum retrial rate $(N\theta)$ is large enough.

1.3 BLOCK GAUSSIAN ELIMINATION

Gaver, Jacobs and Latouche (1984) have applied block gaussian elimination to obtain the stationary distribution for finite level dependent quasi birth-death processes. Bright and Taylor (1995) have suggested applying this method to infinite systems and have given some suggestions on where to truncate the system. They have also provided some probabilistic interpretation to the scheme by identifying the level dependent rate matrices R_i which are generalizations of the matrix geometric rate matrix R in Neuts (1981) and which satisfy the equations

$$A_{i0} + R_i A_{i+1,1} + R_i R_{i+1} A_{i+2,2} = 0.$$

Both of these methods eliminate the rightmost levels first, successively eliminating levels one at a time until the system under consideration consists of the lowest level only.

One problem with this approach is that the decision of where to truncate the system has to be made at the outset, when little information is available. We will reverse this procedure, eliminating the lowest levels first, so that the reduced system is always infinite and, instead of truncating the system, we replace the reduced system with a homogeneous approximation. We can thus gradually increase the level where the homogeneous extension begins until the tail probability above that level is sufficiently small. Since the block Gaussian elimination part of the procedure does not depend on the form of the generator above the levels which are being eliminated, it is not necessary to restart the procedure every time we increase the level where the extension begins.

Since we work in the opposite direction to that of Bright and Taylor (1995), instead of encountering the rate matrices R_i, we encounter the level dependent generalizations $\hat{R}_i$ of the dual $\hat{R}$ of R defined in Hajek (1982). Whereas the elements of R_i represent expected sojourn times in states of level $i+1$ given a chain which starts in level i, the elements of $\hat{R}_i$ represent sojourn times in the states of level $i-1$ given a chain which starts in level i.

Consider the equation $yQ = \alpha$ where Q is a generator or subgenerator. In order to obtain the stationary distribution we must solve this equation with $\alpha = 0$ and Q a generator while obtaining the distribution and moments of the waiting time will require solution of the system with $\alpha \neq 0$ and Q a subgenerator. In applying block Gaussian elimination, we divide the state space into two parts so that the balance equations can be written

$$(y_s, y_t) \begin{bmatrix} Q_s & Q_{st} \\ Q_{ts} & Q_t \end{bmatrix} = (\alpha_s, \alpha_t)$$

or, alternatively

$$y_t Q_t^* = \alpha_t^* \qquad \text{and} \qquad y_s = (\alpha_s - y_t Q_{ts}) Q_s^{-1} \tag{1.3}$$

where

$$Q_t^* = Q_t - Q_{ts} Q_s^{-1} Q_{st} \qquad \text{and} \qquad \alpha_t^* = \alpha_t - \alpha_s Q_s^{-1} Q_{st}. \tag{1.4}$$

For a generator of the form in (1.1), we take $y_s = y_0$, $\alpha_s = \alpha_0$, $y_t = [y_1, y_2, \ldots]$ and $\alpha_t = [\alpha_1, \alpha_2, \ldots]$. Then $Q_s = A_{01}$, $Q_{st} = A_0$, $Q_{ts} = A_2$ and Q_t^* has the form

$$Q_t^* = \begin{bmatrix} A'_{11} & A_{10} & 0 & \ldots \\ A_{22} & A_{21} & A_{20} & \ldots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

where $A'_{11} = A_{11} - A_{12} A_{01}^{-1} A_{00}$. The reduced system $y_t Q_t^* = \alpha_t^*$ can now be solved and the solution substituted into (1.3) to obtain y_s. When combined with the solution for y_t, this yields the entire solution $y = (y_s, y_t)$. We can interpret the reduced system $y_t Q_t^* = \alpha_t^*$ as describing the evolution of the Markov chain on the set of levels $\{1, 2, \ldots\}$ i.e. this is how we would describe the system if we could only view the chain when it was in one of those levels. There are two advantages to using this procedure. First, if Q is a generator (subgenerator), the above calculations do not require subtractions , so the procedure is numerically stable. Second, the matrix Q_t^* of the reduced system is also a generator (subgenerator) for a (level dependent) QBD process. This allows us to apply the process repeatedly, eliminating one level at each step. At the ith step, we obtain a generator (subgenerator) which describes the evolution of the chain on the subset of levels $\{i, i+1, \ldots\}$ and has the form

$$Q_i = \begin{bmatrix} A'_{i1} & A_{i0} & 0 & 0 & \ldots \\ A_{i+1,2} & A_{i+1,1} & A_{i+1,0} & 0 & \ldots \\ 0 & A_{i+2,2} & A_{i+2,1} & A_{i+2,0} & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{1.5}$$

where $A'_{i1} = A_{i1} - A_{i,2}(A'_{i-1,1})^{-1}A_{i-1,0}$. This process is referred to as the reduction phase of the algorithm. If the system is homogeneous above level N, then the reduced system with generator (subgenerator) Q_N can be solved using the methods in Neuts (1983)] to obtain $y_t = [y_N, y_{N+1}, \ldots]$. We can substitute this into (1.3) to obtain $y_s = y_{N-1}$, then substitute $y_t = [y_{N-1}, y_N, \ldots]$ into (1.3) to obtain $y_s = [y_{N-2}, y_{N-1}]$ and so on, expanding the solution by one level at a time until we obtain the entire solution $y = [y_0, y_1 \ldots]$. This process is referred to as the expansion phase of the algorithm. The entire algorithm can be summarized as follows:

Algorithm 1.1

Reduction Phase

$i \leftarrow 0$
$A'_{01} \leftarrow A_{01}$
$\alpha'_0 \leftarrow \alpha_0$
Until $i = N$, do
    $A'_{i1} \leftarrow A_{i1} - A_{i,2}(A'_{i-1,1})^{-1}A_{i-1,0}$
    $\alpha'_i \leftarrow \alpha_i - \alpha'_{i-1}(A'_{i-1,1})^{-1}A_{i-1,0}$
    $i \leftarrow i + 1$
End

Middle Phase

Solve $[y_N, y_{N+1}, \ldots]Q_N = [\alpha'_N, \alpha_{N+1}, \ldots]$ where Q_N is defined by (1.5).

Expansion Phase

$i \leftarrow N-1$

Until $i = 0$, do

$y_i \leftarrow (\alpha'_i - y_{i+1}A_{i+1,2})(A'_{i1})^{-1}$

$i \leftarrow i-1$

End

This algorithm (with $\alpha = 0$) was applied by Boel and Talat (1994) to a block tridiagonal generator with level dependent boundary. The time complexity of the algorithm is clearly linear in N. When the generator is finite with N levels and homogeneous throughout, a similar algorithm referred to as the folding algorithm (Ye and Li (1991)) can be applied with time complexity linear in $log_2 N$.

1.4 The Dual Rate Matrices

If we define the matrices $\hat{R}_i = -A_{i2}(A'_{i-1,1})^{-1}$, where the A'_{i1} are defined as in Algorithm 1.1, the recursion in the expansion phase can be written (replacing y with x) as

$$x_{i-1} = x_i \hat{R}_i.$$

This is similar to the equation $x_{i+1} = x_i R_i$ which is used in Bright and Taylor. The recursion for the A'_{i1} can be expressed in terms of the $\hat{R}_i$ as follows:

$$\hat{R}_i = -A_{i2}(A_{i-1,1} + \hat{R}_{i-1}A_{i-2,0})^{-1} \qquad (1.6).$$

or

$$A_{i+1,2} + \hat{R}_{i+1}A_{i1} + \hat{R}_{i+1}\hat{R}_i A_{i-1,0} = 0.$$

The properties of block Gaussian elimination for generators guarantees that the $\hat{R}_i$ are nonnegative and that the matrix inverted in (1.6)

is nonsingular. The $\hat{R}_i$ are the level dependent generalizations of the dual rate matrix $\hat{R}$ defined in Hajek (1982) and the duals of the rate matrices R_i defined in Bright and Taylor (1995). The $\hat{R}_i$ play the same role in the M/G/1 paradigm as the rate matrices R_i play in the GI/M/1 paradigm. Understanding the role of the duals $\hat{R}_i$ requires two modifications to the discussion of the rate matrix R defined in the first chapter of Neuts (1981): We consider generators of the M/G/1 type instead of the GI/M/1 type and we include level dependence in the generator. The analysis is not significantly different from that presented in Neuts (1981) but we include the modified form for the sake of completeness. We begin with discrete time Markov chains, because the analysis is simpler in that context, and then present the modification required for continuous time chains.

Consider a Markov chain with transition probability matrix of the form

$$P = \begin{bmatrix} B_{01} & B_{02} & B_{03} & B_{04} & \cdots \\ B_{10} & B_{11} & B_{12} & B_{13} & \cdots \\ 0 & B_{20} & B_{21} & B_{22} & \cdots \\ 0 & 0 & B_{30} & B_{31} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

The fundamental property of P is that any transition can jump at most one level at a time to the left. We define the taboo probability ${}_iP^{(n)}_{(i,j)(i-k,\nu)}$ to be the probability that, starting in the state (i,j), the chain reaches $(i-k,\nu)$ at time n without returning to the level i in between. Note that if the system cannot return to level i in between it can also not visit any level above i since it must pass through level i to get to $i-k$ from any level above i. The quantities ${}_i\hat{R}^{(k)}_{j\nu}$ defined by

$${}_i\hat{R}^{(k)}_{j\nu} = \sum_{n=0}^{\infty} {}_iP^{(n)}_{(i,j)(i-k,\nu)}$$

are of basic interest. ${}_i\hat{R}^{(k)}_{j\nu}$ is the expected number of visits to the state $(i-k,\nu)$ before the first return to level i given that the chain starts at

(i,j). The matrix with elements ${}_i\hat{R}(k)_{j\nu}$ will be denoted by $\hat{R}_i^{(k)}$ and we will agree to set $\hat{R}_i^{(0)} = I$, the identity matrix, for all $i \geq 0$. We will supress the superscript for the case $k = 1$ so that $\hat{R}_i$ refers to $\hat{R}_i^{(1)}$ and we refer to $\hat{R}_i$ as the ith dual rate matrix of the Markov chain.. We also denote the stationary distribution of the chain by $x = [x_0, x_1, \ldots]$ where each vector x_i corresponds to level i and consists of scalar elements x_{ij} representing the steady state probability of occupying the state (i,j). The following proposition summarizes the required modifications to the results in Section 1.2 of Neuts (1981).

Proposition 1.2: *If the Markov Chain P is positive recurrent, then for $i \geq 0$*

(i) $\hat{R}_i^{(k+1)} = \hat{R}_i \hat{R}_{i-1} \ldots \hat{R}_{i-k}$.

(ii) $\hat{R}_i = \sum_{k=0}^{\infty} \hat{R}_i^{(k)} B_{i-k,k}$.

(iii) $x_{i-1} = x_i \hat{R}_i$.

(iv) $B_{i0}e = \sum_{k=1}^{i} \hat{R}_i^{(k)} \sum_{\nu=k+1}^{\infty} B_{i-k,\nu} e$.

Proof:

(i). Let $n \geq k+1$. By conditioning on the time r of the last visit to the level $i-k$ and on the state $(i-k,h)$ of that visit, before the chain reaches $(i-k-1,\nu)$ at time n, we obtain, for $n \geq k+1$

$$ {}_iP^{(n)}_{(i,j)(i-k-1,\nu)} = \sum_h \sum_{r=0}^{n} {}_iP^{(r)}_{(i,j)(i-k,h)} \quad {}_iP^{(n-r)}_{(i-k,h)(i-k-1,\nu)}. $$

Summation on n now yields

$$ \begin{aligned} [\hat{R}_i^{(k+1)}]_{j\nu} &= \sum_h \sum_{r=0}^{\infty} {}_iP^{(r)}_{(i,j)(i-k,\nu)} \sum_{n'=0}^{\infty} {}_{i-k}P^{(n)}_{(i-k,h)(i-k-1,\nu)} \\ &= \sum_h [\hat{R}_i^{(k)}]_{jh} [\hat{R}_{i-k}]_{h\nu}. \end{aligned} $$

Thus $\hat{R}_i^{(k+1)} = \hat{R}_i^{(k)} \hat{R}_{i-k}$ and the result follows by induction on k.

(ii). Clearly ${}_iP^{(1)}_{(i,j)(i-1,\nu} = [B_{i0}]_{j\nu}$. For $n \geq 2$ we have

$$ {}_iP^{(n)}_{(i,j)(i-1,\nu)} = \sum_h \sum_{k=1}^{\infty} {}_iP^{(n-1)}_{(i,j)(i-k,h)} [B_{i-k,k}]_{h\nu} $$

by conditioning on the state $(i-k,h)$ from which the state $(i-1,\nu)$ is entered at time n. Summation on n yields the required result.

(iii). By conditioning on the time and the state of the last visit to level i, if there is such a visit, we obtain the relation

$$ P^{(n)}_{(i-1,j)(i-1,j)} = {}_iP^{(n)}_{(i-1,j)(i-1,j)} + \sum_{\nu} \sum_{r=0}^{n} P^{(r)}_{(i-1,j)(i,\nu)} \quad {}_iP^{(n-r)}_{(i,\nu)(i-1,j)} $$

for $n \geq 1$ where the quantities $P^{(n)}_{(i,j)(i,)}$ represent the unrestricted probability the chain is in state (i',j') at time n given that it started at (i,j). We add these equations for n ranging from 1 to N and divide the resulting sums by N. As $N \to \infty$, the left hand side tends to $x_{i+1,j}$ by virtue of the classical ergodic theorem for Markov Chains. Since the sum $\sum_{n=1}^{\infty} P^{(n)}_{(i-1,j)(i-1,j}$ is finite, the first term on the right hand side tends to zero. The second term,

$$ \begin{aligned} T_2 &= \sum_{\nu} \frac{1}{N} \sum_{n=1}^{N} \sum_{r=0}^{n} P^{(r)}_{(i-1,j)(i,\nu)} \quad P^{(n-r)}_{(i,\nu)(i-1,j)} \\ &= \sum_{\nu} \frac{1}{N} \sum_{r=0}^{N} P^{(r)}_{(i-1,j)(i,\nu)} \sum_{n=0}^{N-r} {}_iP^{(n)}_{(i,\nu)(i-1,j)} \end{aligned} $$

tends to $\sum_\nu x_{i\nu} [\hat{R}_i]_{\nu j}$ by an elementary summability argument, since $N^{-1} \sum_{r=0}^{N} P^{(r)}_{(i-1,j)(i,\nu)}$ tends to $x_{i\nu}$ and $\sum_{n=0}^{N} {}_iP^{(n)}_{(i,\nu)(i-1,j)}$ has the limit $i\hat{R}_{j\nu}$.

(iv). Since the chain is positive recurrent, if it is at some time in the state (i,j) and a transition to the left of level i occurs, the chain eventually returns to the state (i,j) with probability 1. Thus we can equate the probability that the chain moves to the left with the probability that the chain moves to the left and eventually crosses back to the level i or above. By conditioning on the state (i_1,j_1) of the first visit (after leaving (i,j)) to level i or above and the time n and state (i_2,j_2) of the last visit to a level below i before crossing back, we obtain

$$\begin{aligned}\sum_{j_0}[B_{i0}]_{jj_0} &= \sum_{i_1\geq i}\sum_{i_2\leq i}\sum_{j_1}\sum_{j_2}\sum_{n=0}^{\infty} \\ &\quad \times {}_iP^{(n)}_{(i,j)(i_2,j_2)}[B_{i_2,i_1-i_2+1}]_{j_2j_1} \\ &= \sum_{j_1}\left[\sum_{k=1}^{i}\hat{R}_i^{(k)}\sum_{\nu=k+1}^{\infty}B_{i-k,\nu}\right]_{jj_1}.\end{aligned}$$

The result follows immediately. □

For continuous parameter Markov processes, we consider infinitesimal generators Q which posess the same basic form as the transition probability matrix P except that the diagonal elements are negative and each row of the generator sums to zero instead of one. If Q is positive recurrent then there exists a positive vector x satisfying $xQ = 0$ and $xe = 1$. This balance equation can be rewritten

$$x'_i = \sum_{k=0}^{i+1} x'_k B'_{k,i-k+1}$$

where $x'_i = hx_i\Delta_i$, $B'_{ij} = \delta_{j1}I + \Delta_i^{-1}B_{ij}$, $\Delta_i = -diag(B_{i1})$ and h is any

real number. $xQ = 0$ is equivalent to the equation $x'P = x'$ where

$$P = \begin{bmatrix} B'_{01} & B'_{02} & B'_{03} & B'_{04} & \cdots \\ B'_{10} & B'_{11} & B'_{12} & B_{13} & \cdots \\ 0 & B''20 & B'_{21} & B'_{22} & \cdots \\ 0 & 0 & B'_{30} & B'_{31} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

is the transition matrix for the chain imbedded at each successive transition of the continuous time chain with generator Q. If h is chosen such that $xe = 1$ then x represents the stationary distribution for that chain. Ler $\hat{R}'_i$ denote the ith dual rate matrix of this chain and define $\hat{R}_i$ by

$$\hat{R}_i = \Delta_i \hat{R}'_i \Delta_{i-1}^{-1}.$$

We refer to $\hat{R}_i$ as the ith dual rate matrix of the continuous time chain Q and interpret its elements as follows: We can write

$$[\hat{R}_i]_{j\nu} = \left(\frac{[\hat{R}'_i]_{j\nu}}{[\Delta_{i-1}]_{\nu\nu}} \right) [\Delta_i]_{jj}$$

where $[\Delta_i]_{jj}$ represents the inverse of the expected sojourn time in the state (i, j). Thus $[\hat{R}_i]_{j\nu}$ represents the time spent in state $(i-1, \nu)$ before the first return to level i, measured in units of duration $[\Delta_i]_{jj}^{-1}$ which is the mean sojourn time in the state (i, j).

The continuous time analog of Proposition 1.2 follows from that proposition and the definition $\hat{R}_i = \Delta_i \hat{R}'_i \Delta_{i-1}^{-1}$:

Proposition 1.3: *If the Markov Chain Q is positive recurrent, then for $i \geq 0$*

(i) $\hat{R}_i^{(k+1)} = \hat{R}_i \hat{R}_{i-1} \ldots \hat{R}_{i-k}$.

(ii) $0 = \sum_{k=0}^{\infty} \hat{R}_i^{(k)} B_{i-k,k}$.

(iii) $x_{i-1} = x_i \hat{R}_i$.

(iv) $B_{i0}e = \sum_{k=1}^{i} \hat{R}_i^{(k)} \sum_{\nu=k+1}^{\infty} B_{i-k,\nu}e$. □

The recursion formula *(iii)*for the x_i is clearly familiar and equation (1.6) can be obtained from *(ii)*if we set $B_{i0} = A_{i2}$, $B_{i1} = A_{i1}$ and $B_{i2} = A_{i0}$.

1.5 STATIONARY DISTRIBUTION

In order to approximate the stationary distribution of the exact queue, we solve for the stationary distribution of the first approximation, in which the server is never idle at levels above N, and increase the value of N until the probability of states above that level becomes insigificant. In order to obtain the stationary distribution for the approximate queue, we must solve the equation $xQ = 0$ where Q is a generator. We can do this by executing Algorithm 1 with $\alpha = 0$, and setting the stationary distribution x equal to the result y. The reduction phase eliminates levels $0, 1, \ldots, N-1$ in that order and the middle phase finds the component x_N (up to a constant factor) by applying methods in Neuts (1981). The expansion phase then finds $x_{N-1}, x_{N-2}, \ldots, x_0$ in that order. If we were to apply this algorithm directly to the MAP/PH/1 retrial queue, the time complexity of the reduction phase would be $O(Nm^3n^3)$ where m and n are the dimensions of the arrival and service process representations respectively. We can reduce this to $O(Nm^3n^2)$ by taking advantage of the sparsity of A_0 and the A_{i2}.

The interpretation of the matrices $\hat{R}_i$ makes it clear that they must have the form

$$\hat{R}_i = \begin{bmatrix} M_i & \delta_i^T \\ 0 & 0 \end{bmatrix}$$

since any chain started in a state (i, j) in which the server is busy cannot

reach level $i-1$ without first visiting some state at level i in which the server is idle. This can also be shown by induction on the level i. If we combine the definition $\hat{R}_i = -A_{i2}(A'_{i-1,1})^{-1}$ with the recursion for the A'_{i1} given in Algorithm 1, we obtain, after substituting from (1.2), the following recursion for the M_i and δ_i^T:

$$\delta_0 = 0$$

$$M_0 = 0$$

$$\delta_{i+1}^T = (i+1)\theta[\nu^T + JL_i\gamma_i^T]$$

$$M_{i+1} = (i+1)\theta JL_i \tag{1.7}$$

where

$$\nu^T = -(I \otimes \beta^T)(I \otimes S + D_0 \otimes I)^{-1}$$

$$J = \nu^T(I \otimes S^0)$$

$$\gamma_i^T = -[D_1 \otimes \beta^T + \delta_i^T(D_1 \otimes I)](I \otimes S + D_0 \otimes I)^{-1}$$

$$L_i = -[D_0 - i\theta I + \gamma_i^T(I \otimes S^0)]^{-1}. \tag{1.8}$$

The components $x_i = [x_i^0, x_i^1]$ $(x_i^0 \in \Re^m)$ $i = 0, 1, \ldots, N-1$ of the stationary distribution are given by

$$x_i^0 = x_N^0 M_N M_{N-1} \ldots M_{i+1}$$

$$x_i^1 = x_{i+1}^0 \delta_{i+1}^T. \tag{1.9}$$

The generator for the approximate queue has the form

$$Q = \begin{bmatrix} A_{0,1} & A_0 & 0 & \cdots & & & & \\ A_{1,2} & A_{1,1} & A_0 & 0 & \cdots & & & \\ 0 & \ddots & \ddots & \ddots & & & & \\ \vdots & & A_{N2} & A_{N1} & A_{N0} & & & \\ & & & A_{N+1,2} & A'_1 & A'_0 & & \\ & & & & A'_2 & A'_1 & A'_0 & \\ & & & & & \ddots & \ddots & \ddots \end{bmatrix}$$

where A_0,A_{i1} and A_{i2} are as in (1.2) for $i = 0, 1, \ldots, N$, $A'_0 = D_1 \otimes I$, $A'_1 = I \otimes S + D_0 \otimes I$, $A'_2 = I \otimes S^0\beta^T$,

$$A_{N0} = \begin{bmatrix} 0 \\ D_1 \otimes I \end{bmatrix} \qquad \text{and} \qquad A_{N+1,2} = [\,0 \quad I \otimes S^0\beta^T\,].$$

If we view this chain only when it is in level N, we can describe the evolution by the generator

$$A''_{N1} = \begin{bmatrix} D_0 - N\theta I & D_1 \otimes \beta^T + \delta_N^T(D_1 \otimes I) \\ I \otimes S^0 & I \otimes S + D_0 \otimes I + R(I \otimes S^0\beta^T) \end{bmatrix}$$

where R is the minimal nonnegative solution to the matrix quadratic equation

$$D_1 \otimes I + R(I \otimes S + D_0 \otimes I) + R^2(I \otimes S^0\beta^T) = 0. \tag{1.10}$$

Methods for solving this equation are discussed in Neuts (1981). We might think of A''_{N1} as if it was obtained from Q_N (see (1.5)) by truncating at some level above N, and by applying block Gaussian elimination as in Gaver, Jacobs and Latouche (1984), (eliminating the topmost level first and then successively lower levels, one at a time) until only level N remained. In fact, since in practice we must terminate the algorithm which obtains R after a finite number of steps, if we use the algorithm $X \leftarrow 0$; Loop$\{X \leftarrow -A_0(A_1 + XA_2)^{-1}\}$ to obtain R instead

of the usual one ($X \leftarrow 0$; Loop$\{X \leftarrow -(A_0 + X^2 A_2)A_1^{-1}\}$), then this characterization is correct. Neuts (1981) has suggested that the former algorithm, while it usually requires less iterations than the latter to converge to R, often requires more computation time because of the effort required to invert the matrices. This description of A''_{N1} may not be exactly correct if we use the latter algorithm however the connection to block Gaussian elimination seems clear.

We can now obtain x_N by solving $x_N A''_{N1} = 0$ or by solving

$$x_N^0[D_0 - N\theta I - (D_1 \otimes \beta^T + \delta_N^T(D_1 \otimes I))(I \otimes S + D_0 \otimes I + R(I \otimes S^0\beta^T))^{-1}(I \otimes S^0)] = 0 \tag{1.11}$$

and then setting

$$x_N^1 = -x_N^0[D_1 \otimes \beta^T + \delta_N^T(D_1 \otimes I)][I \otimes S + D_0 \otimes I + R(I \otimes S^0\beta^T)]^{-1} \tag{1.12}$$

and $x_{N+j} = x_N^1 R^j$ for $j = 1, 2, \ldots$. Of course (1.11) must be solved subject to a normalization condition. Define η_i and ξ_i according to

$$\eta_0 = 0 \quad ; \quad \eta_i = M_i(\eta_{i-1} + e)$$

$$\xi_0 = 0 \quad ; \quad \xi_i = \delta_i^T e + M_i \xi_{i-1}. \tag{1.13}$$

Then $x_N^0 \eta_N = \sum_{i=0}^{N-1} x_i^0 e$, $x_N^0 \xi_N = \sum_{i=0}^{N-1} x_i^1 e$ and the normalization condition $xe = 1$ can be written

$$x_N^0[\eta_N + \xi_N + e - (D_1 \otimes \beta^T + \delta_N^T(D_1 \otimes I))(I \otimes S + D_0 \otimes I + R(I \otimes S^0\beta^T))^{-1}(I - R)^{-1}e] = 1 \tag{1.14}$$

Once we have scaled x to satisfy (1.14), we will be interested in the tail probability

$$P_N = \sum_{i=N}^{\infty} x_i e = 1 - x_N^0(\eta_N + \xi_N) \tag{1.15}$$

above level $N-1$. We propose to increase N until $P_N < \epsilon$ for some small $\epsilon \in (0,1)$ and then replace x with the finite vector

$$x_f = (1 - P_N + x_N e)^{-1}[x_0, x_1, \ldots, x_N]. \tag{1.16}$$

We summarize this procedure as follows:

Algorithm 1.2

$M_0, \delta_0^T, \eta_0, \xi_0 \leftarrow 0$
$N \leftarrow 0$
Calculate R from (1.10)
Until $P_N < \epsilon$, do
    For $i = N$ to $N+n$, Calculate M_i,δ_i^T, η_i and ξ_i from (1.7), (1.8) and (1.12)
    $N \leftarrow N+n$
    Calculate x_N^0 from (1.11) and (1.14)
    Calculate P_N from (1.15)
    Choose a new integer step n
End Loop
Calculate x_N^1 via (1.12)
Calculate $x_{N-1}, x_{N-2}, \ldots, x_0$ from (1.9)
Approximate x with x_f defined in (1.16)

In order to illustrate the methods developed in this chapter, we consider a particular case of a MAP/PH/1 retrial queue and apply the methods developed to the analysis of this queue. The example we consider has retrial rate $\theta = 2$, service time distribution (β, S) with

$$\beta^T = [1 \quad 0 \quad 0] \qquad S = \begin{bmatrix} -6 & 6 & 0 \\ 0 & -6 & 6 \\ 0 & 0 & -6 \end{bmatrix}$$

and arrival process (D_0, D_1) where

$$D_0 = \begin{bmatrix} -3 & 2 & 0 \\ 0 & -2 & 1.8 \\ 0 & 0 & -4 \end{bmatrix} \qquad D_1 = \begin{bmatrix} .5 & 0 & .5 \\ .2 & 0 & 0 \\ 3 & .4 & .6 \end{bmatrix}.$$

Figure 1.1 shows the estimate of the tail probability P_N based on the first approximation as well as a similar estimate based on the second approximation with bounded retrial rate. From the graph, it is clear that the tail probability of the second approximation dominates. In the next section, we prove that this estimate is, in fact an upper bound on the true tail probability. We chose to truncate the generator above level $N = 30$ which yields an estimate for P_N of approximately 10^{-5}. The upper bound (from the second approximation) on the tail probability above level $N = 30$ is 1.15×10^{-5}.

Figure 1.2 compares the cumulative distribution of the number of customers in the system to the same distribution for a nonretrial queue (i.e. with infinite buffer) with the same arrival and service processes. As expected, the number of customers is smaller in the nonretrial queue since, unlike the retrial queue, the server is never idle when the system is nonempty.

Figure 1.1: Estimates of Tail Probability for Queue

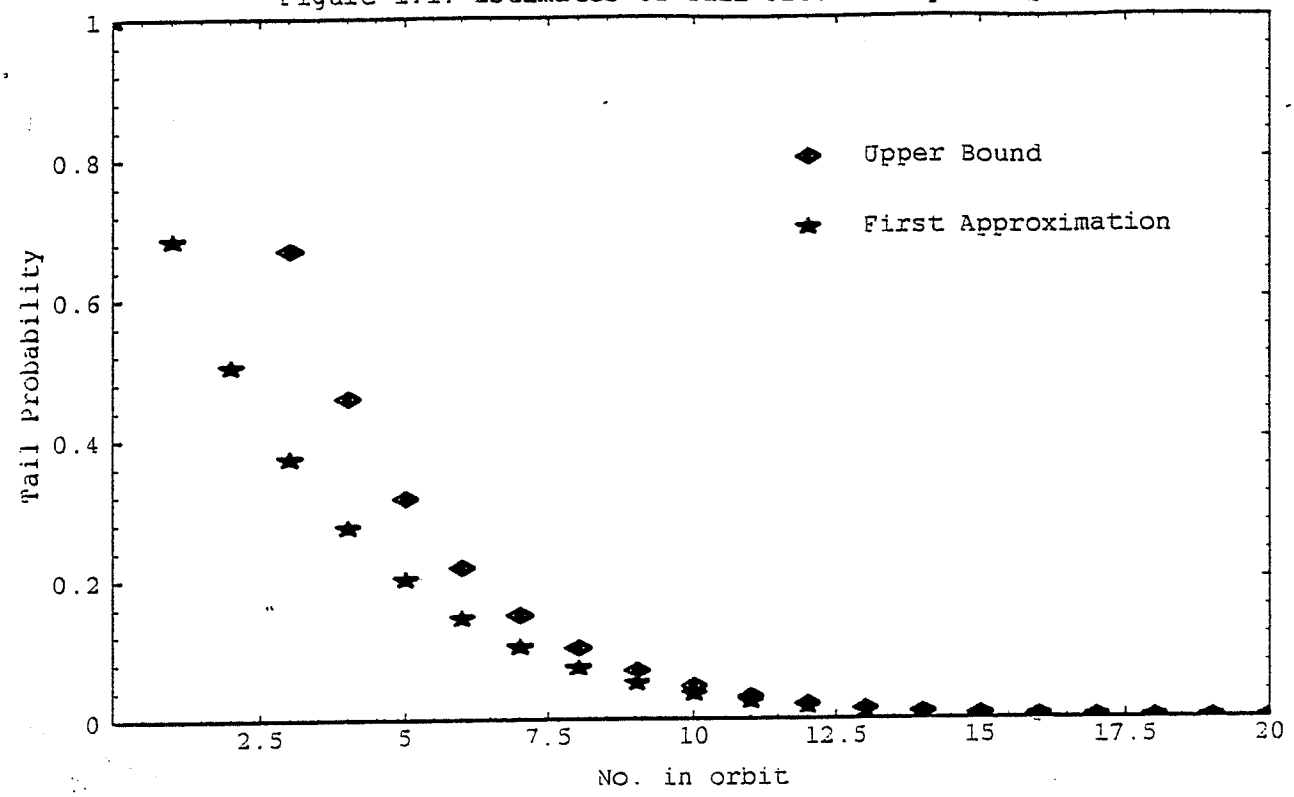
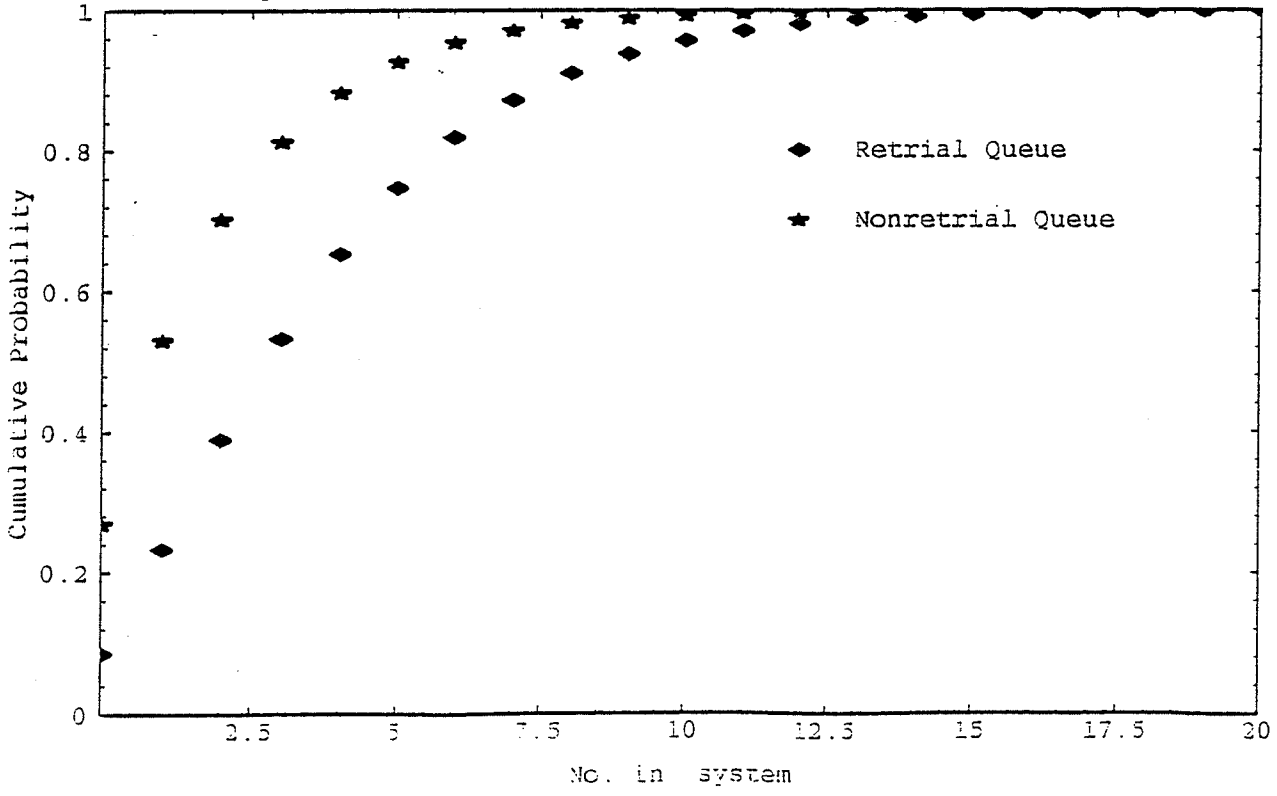


Figure 1.2: CDF of No. in System (Retrial and Nonretrial)



1.6 BOUND ON TAIL PROBABILITY

We refer to the queue with the homogeneous extension used by Neuts and Rao (1990) (with bounded retrial rate) as the second approximation. Unlike the first approximation used in the previous section, the rate matrix R associated with this extension depends on the level N at which it begins and so must be recalculated if this level is increased. However, it is a pessimistic approximation and so, under certain conditions, can give us an upper bound on the tail probability above level N in the exact queue. The first approximation can be used, via Algorithm 1.2, to determine N without repeated calculations of R, and the bound from the second approximation can be used afterward as a check to guarantee that the tail probability of the exact queue is sufficiently small.

In this section, we define a class of phase type distributions. The tail probability of any MAP/PH/1 retrial queue with a service time distribution from this class is bounded above by the tail probability of the second approximation. In order to prove this, we make use of theorems in Massey (1987) to show that the second approximation stochastically dominates the exact queue. Bright and Taylor (1995) have applied this method, considering service time distributions in which service completion can occur from any stage of service.

We begin by presenting some theoretical results from Massey (1987). For a set E, $\leq_E$ is a partial order on E if

(i) $s \leq_E s \quad \forall s \in E$
(ii) $s \leq_E t$ and $t \leq_E s$ imply $s = t$
(iii) $s \leq_E t$ and $t \leq_E u$ imply $s \leq_E u$.

Consider a Markov process on a state space E upon which is defined a partial order $\leq_E$. For any subset Γ of E we define $\Gamma^{\uparrow} = \{y | y \geq_E x$ for some $x \in \Gamma\}$. A subset Γ is said to be an increasing set if $\Gamma = \Gamma^{\uparrow}$.

The partial order on E induces a partial order on probability measures defined on E: If p and q are probaility measures , we say $p \leq_E q$ if $p(\Gamma) \leq q(\Gamma)$ for all increasing sets $\Gamma \subseteq E$. We can also define a partial order on generators for Markov processes on E as follows: If A and B are bounded linear operators on $\ell_1(E)$ we say that $A \leq_E B$ provided $Ae_\Gamma \leq_E Be_\Gamma$ for all increasing sets $\Gamma \subseteq E$ where e_Γ denotes the indicator function for the set Γ. Proposition 3.1 in Massey (1987) proves that these are, in fact, partial orders. A generator A for a Markov process on E is said to be strongly monotone if for all probability vectors p and q in $\ell_1(E)$, $p \leq_E q \Longrightarrow p\ exp(At) \leq_E q\ exp(At)$ for all $t > 0$. A Markov process with generator B is said to stochastically dominate the process with generator A if $p \leq_E q \Longrightarrow p\ exp(At) \leq_E q\ exp(Bt)$ for all $t > 0$.

According to Theorem 3.4 in Massey (1987), if A and B are bounded, $A \leq_E B$ and one of A or B is monotone, then the process with generator B stochastically dominates the process with generator A. Brandt and Last (1993) have removed the requirement of bounded generators from Theorem 5.3 in Masssey and their version can be used to generalize Theorem 3.4 to the unbounded case as well. We will also make use of Theorem 4.1 in Massey which can be similarly generalized to the unbounded case.

A phase type distribution is the distribution of the time to absorption in a finite absorbing Markov chain. For distributions in class 1 defined below, the associated Markov chains are monotonic by virtue of Theorem 4.1 in Massey (1987). This monotonicity is inherited by the Markov chain representing the MAP/PH/1 retrial queue with the same service time distribution. Theorem 3.4 in Massey can then be used to show that the second approximation stochastically dominates the exact queue and so has a larger tail probability.

Definition 1.1: We say that a phase type distribution with repre-

sentation (β, S) on the state space σ_s belongs to Class 1 if there exists a partial order (denoted by $\leq_s$) on the state space $\sigma_s^* = \{0\} \cup \sigma_s$ ($\{0\}$ represents the absorbing state in which service is completed), a set of nonnegative scalars $\{\gamma_k | k = 1, \ldots, K_s\}$ and a set of functions $\{f_k : \sigma_s^* \to \Re | k = 1, \ldots, K_s\}$ such that

(1). $[\beta]_\ell > 0 \quad \Longrightarrow \ell \in max(\sigma_s^*)$

(2).

$$S^* = \begin{bmatrix} S^0 & S \\ 0 & 0 \end{bmatrix} = \sum_{k=1}^{K_s} \gamma_k (\Phi(f_k) - I)$$

where for any function $f : \sigma_s^* \to \sigma_s^*$, $\Phi(f)$ is the matrix defined by

$$[\Phi(f)]_{ij} = \begin{cases} 1 & j = f(i) \\ 0 & otherwise \end{cases}$$

(3). The fuctions $\{f_k\}$ are monotone:

$$i \leq_s j \Longrightarrow f_k(i) \leq_s f_k(j) \quad \square$$

This class of distributions has the following closure property:

Proposition 1.4: *Class 1 is closed under convolution.*

Proof: Let (α, S) and (β, T) belong to class 1 and denote their convolution by (η, H) where

$$\eta^T = (\alpha^T, 0) \qquad H = \begin{bmatrix} S & S^0 \beta^T \\ 0 & T \end{bmatrix}.$$

We denote the state space associated with this representation by σ_h^*:

$$\sigma_h^* = \{(0, i) | i \in \sigma_t^*\} \cup \{(1, j) | j \in \sigma_s\}$$

$(\sigma_t^* = \sigma_t \cup \{0\})$. Let $\leq_s$ and $\leq_t$ denote the partial orders on σ_s and σ_t^* respectively and define the partial order $\leq_h$ on σ_h^* according to the following:

(i) $(0,0) \leq_h (0,i) \leq_h (1,j) \qquad \forall i \in \sigma_t^*, \quad j \in \sigma_s.$
(ii) $(0,i) \leq_h (0,j) \quad iff \quad i \leq_t j.$
(iii) $(1,i) \leq_h (1,j) \quad iff \quad i \leq_s j.$

Let $\{f_k | k = 1, \ldots, K_s\}$ and $\{g_k | k = 1, \ldots, K_t\}$ be the monotone fuctions (defined on σ_s and σ_t respectively) in properties (2) and (3) of the definition of Class 1. Define the functions $\{h_{k\ell} : \sigma_h \to \sigma_h^* | k = 1, \ldots, K_s + K_t, \quad \ell \in \sigma_t, \quad [\beta]_\ell > 0\}$ according to:

$$h_{k\ell}(1,j) = \begin{cases} (0,\ell) & f_k(j) = 0 \quad and \quad k \leq K_s \\ (1, f_k(j)) & f_k(j) \neq 0 \quad and \quad k \leq K_s \\ (1,j) & otherwise \end{cases}$$

$$h_{k\ell}(0,j) = \begin{cases} (0, g_{k-K_s}(j)) & k > K_s \\ (0,j) & otherwise \end{cases}.$$

Let γ_k and δ_k denote the scalars associated with f_k and g_k respectively in property (2) of the definition of class 1. With each function $h_{k\ell}$ we associate the scalar

$$\epsilon_{k\ell} = \begin{cases} \gamma_k [\beta]_\ell & k \leq K_s \\ \delta_k [\beta]_\ell & k > K_s \end{cases}$$

Then we have

$$H^* = \begin{bmatrix} H^0 & H \\ 0 & 0 \end{bmatrix} = \sum_{k=1}^{K_s+K_t} \sum_{\ell \in \sigma_t} \epsilon_{k\ell} (\Phi(h_{k\ell}) - I).$$

It remains to show that the $h_{k\ell}$ are monotonic with respect to $\leq_h$. We consider separately the three cases ((i), (ii) and (iii above) which define the partial order $\leq_h$:

(i). $h_{k\ell}(0,i) \leq_h (0,\ell) \leq_h h_{k\ell}(1,j)$.

(ii). Let $i \leq_t j$. If $k \leq K_s$, then $h_{k\ell}(0,i) = (0,i) \leq_h (0,j) = h_{k\ell}(0,j)$. If $k > K_s$, then $h_{k\ell}(0,i) = (0, g_{k-K_s}(i)) \leq_h (0, g_{k-K_s}(j)) = h_{k\ell}(0,j)$.

(iii). Let $i \leq_s j$. If $k \leq K_s$ and $f_k(i) \neq 0$ then $h_{k\ell}(1,i) = (1, f_k(i)) \leq_h (1, f_k(j)) = h_{k\ell}(1,j)$. If $k \leq K_s$ and $f_k(i) = 0$, then $h_{k\ell}(1,i) = (0,\ell) \leq_h h_{k\ell}(1,j)$. If $k > K_s$, then $h_{k\ell}(1,i) = (1,i) \leq (1,j) = h_{k\ell}(1,j)$.

Thus $i \leq_h j \Longrightarrow f_{k\ell}(i) \leq_h f_{k\ell}(j)$ and the proposition is proved. □

It can easily be shown that hyperexponential distributions are in class 1. A Coxian distribution is a member of this class if the probability of service completion after each stage increases monotonically with the maximum number of stages ahead (i.e. the number of stages left to visit for a customer who visits every stage).

Asmussen and Nerman (1991) have applied the EM algorithm, a maximum likelihood approach, to fitting arbitrary phase type distributions to empirical data. An implementation under the name EMPHT by Häggström, Asmussen and Nerman (1991) allows the user to specify which transitions between states of the underlying process are allowed. This implementation could thus be used to fit phase type distributions which are convolutions of hyperexponentials and thus members of class 1. It has been demonstrated empirically (see Whitt (1984)) that for queues with service time distriutions having coefficients of variation smaller than 1.0, it is often sufficient to match the first two moments of the distribution to obtain a good approximation. We can always match the first two moments of a distribution with c.v. $c < 1$ to a mixture of two Erlang distributions of order n and $n+1$ and common rate parameter where $1/(n+1) \leq c \leq 1/n$. These distributions are members of class 1. Hyperexponential distributions could be used (see Altiok (1985)) to

match the first three moments for the case $c > 1.0$. If empirical service time moments can be matched to a distribution in this class, it would be possible to obtain bounds on tail probabilities for MAP/PH/1 retrial queues with the chosen service time distribution.

The following proposition applies a theorem (4.1) from Massey (1987) to show that a MAP/PH/1 retrial queue is strongly monotone if the service time distribution is a member of class 1.

Proposition 1.5: *If the service time distribution (β, S) for a MAP/PH/1 retrial queue is in class 1, then the queue is strongly monotone with respect to the following partial order:*

$$(r,s,t) \leq (r',s',t') \quad iff \quad s = s', \quad r \leq r \quad and \quad (r < r' \quad or \quad t \leq_s t') \tag{1.17}$$

where r,s and t represent the number of customers in orbit, phase of arrival and phase of service ($t = 0$ if server is idle) respectively.

Proof: Let γ_k and f_k be the scalars and functions referred to in the defiition of class 1 and define the following functions and scalars:

$$F_k(r,s,t) = (r,s,f_k(t)) \qquad k = 1,\ldots,K_s$$

$$d^0_{ij}(r,s,t) = \begin{cases} (r,j,t) & s = i \\ (r,s,t) & otherwise \end{cases} \qquad i,j = 1,\ldots,m$$

$$\delta^0_{ij} = [D_0]_{ij}$$

$$d^1_{ij\ell}(r,s,t) = \begin{cases} (r,j,\ell) & s = i, \quad t = 0 \quad and \quad [\beta]_\ell > 0 \\ (r+1,j,t) & s = i, \quad t \neq 0 \quad and \quad [\beta]_\ell > 0 \\ (r,s,t) & otherwise \end{cases}$$

$$\delta^1_{ij\ell} = [D_1]_{ij}[\beta]_\ell$$

$$\Theta_{j\ell}(r,s,t) = \begin{cases} (r-1,s,\ell) & t = 0 \quad and \quad r \geq j \\ (r,s,t) & otherwise \end{cases}$$

$$\theta_\ell = \theta[\beta]_\ell. \tag{1.18}$$

The generator for the queue can now be expressed as

$$\begin{aligned} Q = &\sum_{k=1}^{K_s} \gamma_k(\Phi(F_k) - I) + \sum_{i \neq j \in \{1,\ldots,m\}} \delta_{ij}^0(\Phi(d_{ij}^0) - I) \\ &+ \sum_{i,j=1}^{m} \delta_{ij\ell}^1(\Phi(d_{ij\ell}^1) - I) + \sum_{j=0}^{\infty} \sum_{\{\ell | [\beta]_\ell > 0\}} \theta_{j\ell}(\Phi(\Theta_{j\ell}) - I). \end{aligned} \tag{1.19}$$

It is easy to see that d_{ij}^0 is monotonic with respect to the given partial order. F_k is monotone by virtue of property (3) of the definition of class 1 and $\Theta_{j\ell}$ and $d_{ij\ell}^1$ are monotone by virtue of property (1) of that definition. The queue is thus monotone with respect to the given partial order by theorem 4.1 in Massey (1987) and an extension to unbounded generators obtained from Brandt and Last (1993). □

Note, as a corollary to Proposition 1.5, that the second approximation for the queue is also monotone. We can obtain this approximation by replacing the infinite upper limit in the expression for Q with N. We can now compare the exact queue to the second approximation.

Proposition 1.6: *If the service time distribution for a MAP/PH/1 retrial queue is in class 1 then the second approximation (i.e. where the total rate of retrial attempts from orbit is bounded by $N\theta$) stochastically dominates the exact queue with respect to the partial order (1.17).*

Proof: Let Q and Q' denote respectively the generators for the exact queue and the approximation. We can obtain Q' from the expression (19) for Q if we replace $\Theta_{j\ell}$ with

$$\Theta'_{j\ell}(r,s,t) = \begin{cases} (r-1,s,\ell) & t = 0, \quad j \leq r \quad and \quad j \leq N \\ (r,s,t) & otherwise \end{cases}. \tag{1.20}$$

Since $\Theta_{j\ell} \leq \Theta'_{j,l}$, we can show that $Q \leq Q'$ by an argument similar to the proof of Theorem 4.2 in Massey (1987). The result is proved by applying Proposition 1.5 together with Theorem 3.4 in Massey . □

This stochastic dominance clearly guarantees that the tail probability of the second approximation is larger than that of the exact queue. Since the stochastic dominance holds for all time, it may be possible to obtain similar results concerning other performance criteria, however we do not consider these here.

1.7 DISTRIBUTION OF THE NUMBER OF RETRIALS

In the following sections, we assume the following approximate form for the generator of the queue:

$$Q = \begin{bmatrix} A_{0,1} & A_0 & 0 & \dots & \\ A_{12} & A_{11} & A_0 & 0 & \dots \\ 0 & \ddots & \ddots & \ddots & \\ \vdots & & \ddots & \ddots & A_0 \\ & & & A_{N2} & B_{N1} \end{bmatrix}$$

where

$$B_{N1} = A_{N1} + \begin{bmatrix} 0 & 0 \\ 0 & R(I \otimes S^0 \beta^T) \end{bmatrix}$$

and R is the solution of (1.10). This is the generator which describes the evolution of the first approximation on the set of levels $\{0, 1, \dots, N\}$.

For retrial queues it is natural to measure the waiting time not only in absolute units but also by the number of retrials performed by an arbitrary customer. It is an important quantity in itself because it determines the additional load on control devices for some systems. Let n denote the number of retrials performed by a randomly chosen customer before entering into service. We obtain the probability of

immediate service $P(n=0)$ as follows. Let Q_s be the matrix constaining the elements of Q which correspond to transitions in which a customer arrives and, finding the server idle, begins his service immediately. Q_s is block diagonal with blocks

$$A_s = \begin{bmatrix} 0 & D_1 \otimes \beta^T \\ 0 & 0 \end{bmatrix}.$$

Similarily, let Q_t be the matrix containing the elements of Q corresponding to transitions where an arriving customer enters the orbit. Q_t is obtained from Q by deleting the diagonal and subdiagonal blocks. Also define $Q_0 = Q - Q_s - Q_t$. We can find the probability that an arbitrary customer enters service immediately upon arrival by considering an absorbing Markov chain with the following generator:

$$Q' = \begin{bmatrix} Q_0 & Q_s & Q_t \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

The stationary distribution of states after an arbitrary arrival is given by

$$x_a = \frac{x(Q_s + Q_t)}{x(Q_s + Q_t)e}.$$

This is the stationary vector of the transition probability matrix $-Q_0^{-1}(Q_s + Q_t)$ which evolves the system from one arrival to the next. The probability of immediate service is just the probability of eventual absorption into the "s group "of states given that we start the chain at an arbitrary arrival. This probability is given by

$$P(n=0) = -x_a Q_0^{-1} Q_s e = \frac{xQ_s e}{x(Q_s + Q_t)e}. \tag{1.21}$$

The stationary distribution of states after an arrival which enters the orbit is given by

$$x_t = \frac{xQ_t}{xQ_t e}.$$

This is the stationary vector of the transition probability matrix $-(Q_0+Q_s)^{-1}Q_t$ which evolves the system from one orbit entry to the next. Clearly x_a and x_t are easy to calculate (in $O(Nm^2n^2)$ time) once x is known.

The conditional distribution of the number n of retrials performed by a customer who enters the orbit is given by the following:

Proposition 1.7: *The conditional probability distribution for the number of retrials n per customer in an M/PH/1 retrial queue is given by:*

$$P(n = k|n > 0) = x_t \left[P_b\theta(\theta I - Q)^{-1}\right]^k (I - P_b)e$$

Proof: Let $p^k \in \Re^\infty$ $(k = 0, 1, 2, \ldots)$ be such that $p^k_{(i,j)}$ is the probability that a randomly chosen customer makes at least k retrials (not including the request for service upon arrival) before being served and that the system is in the state (i, j) immediately after the $k-1$st retrial. We consider the request for service made immediately upon arrival to be the zeroth retrial. Similarly, let $q^k \in \Re^\infty$ be such that $q^k_{(i,j)}$ is the probability that the customer makes at least k retrials and that the system is in the state (i, j) immediately before the kth retrial. If the system is in equilibrium , then $q^0 = x_t$ (the stationary distribution after arrivals) and $p^1 = x_t P_b$. By conditioning on the time s between the $k-1$st and kth retrial, (recall s is exponentially distributed with mean θ^{-1}) we obtain

$$q^k = p^k \int_0^\infty \theta\ exp(-\theta s)\ exp(Qs)ds = p^k\theta(\theta I - Q)^{-1}.$$

Combining this with the relation $p^k = q^{k-1}P_b$ yields, by induction on k,

$$q^k = x_t \left[P_b\theta(\theta I - Q)^{-1}\right]^k.$$

After each retrial, the customer enters service if the server is idle so that $P(n = k|n > 0) = q^k(I - P_b)$. □

The operator $\theta(\theta I - Q)^{-1}$ is a stochastic matrix whose elements are the transition probabilities for the Markov chain embedded at successive retrials. Premultiplying by P_b creates a substochastic matrix which is identical to $\theta(\theta I - Q)^{-1}$ except for the rows corresponding to idle states (where the server is idle) which become zero. This allows the system to escape from these states to an absorbing state which we can add to the state space and which corresponds to the customer in question having already left the orbit. The distribution of the number of retrials is of phase type with infinite dimensional representation $(x, P_b\theta(\theta I - Q)^{-1})$. We can calculate the action of the operator $\theta(\theta I - Q)^{-1}$ on a vector recursively if we can solve equations of the form $y(\theta I - Q) = \alpha$. This can be done by making the substitution $D_0 \leftarrow D_0 - \theta I$ and then applying the following :

Algorithm 1.3

$M_0, \delta_0^T \leftarrow 0$
For $i = 1, \ldots, N$, do : Calculate L_i γ_i^T, M_i and δ_i^T from (1.7) and (1.8)
Calculate R from (1.10)

$\alpha_0' \leftarrow \alpha_0$
For $i = 1, \ldots, N$, do : $\alpha_i' \leftarrow \alpha_i + \alpha_{i-1}' Y_i$
$y_N \leftarrow \alpha_N' X_N$
For $i = N, N-1, \ldots, 0$ do : $y_i \leftarrow \alpha_i' X_i + y_{i+1}\hat{R}_{i+1}$

where

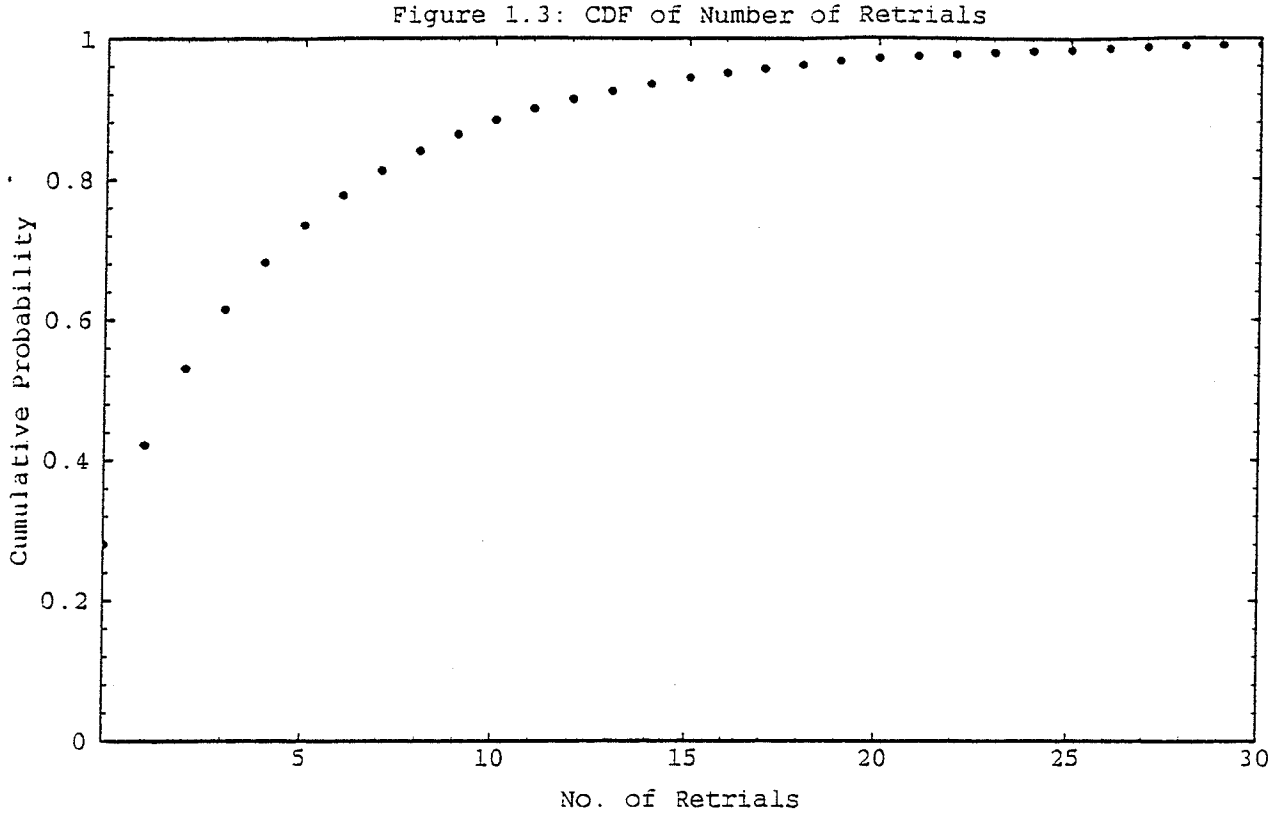
$$Y_i = (-A'_{i1})^{-1} A_0 = \begin{bmatrix} 0 & L_i \gamma_i^T (D_1 \otimes I) \\ 0 & T_i[I + (I \otimes S^0) L_i \gamma_i^T] \end{bmatrix}$$

$$X_i = (-A'_{i1})^{-1} = \begin{bmatrix} L_i & L_i \gamma_i^T \\ T_i(I \otimes S^0) L_i & T_i[I + (I \otimes S^0) L_i \gamma_i] \end{bmatrix}$$

$$\hat{R}_i = \begin{bmatrix} M_i & \delta_i^T \\ 0 & 0 \end{bmatrix}.$$

Recall that, for this particular application we have made the substitution $D_0 \leftarrow D_0 - \theta I$ before executing Algorithm 1.3 so that R ,L_i, γ_i^T, M_i and δ_i^T are different than they were in executing Algorithm 1.2. The elements of $\hat{R}_i$ must now be interpreted as expected sojourn times in level $i-1$ given a chain starting in level i before the first return to level i or the next retrial of our tagged customer. The interpretation of R is also similarly altered. Clearly we must apply this algorithm M times to calculate $P(n = k)$ for $k = 1, 2, \ldots, M$. However, the first part of Algorithm 1.3, which has a time complexity of $O(Nm^3n^2)$, only needs to be executed once. The time required to get all M probabilities is thus $O(Nm^3n^2) + O(NMm^2n^2)$. Figure 1.3 shows the cumulative distribution of the number of retrials for the example in Section 1.5. The distribution in the figure is the unconditional one: $P(n \leq k)$ and not $P(n \leq k | n > 0)$.

Figure 1.3: CDF of Number of Retrials



1.8 WAITING TIME DISTRIBUTION

If a customer does not enter service immediately, his waiting time in the orbit has a phase type distribution with representation $(x_t, Q - \theta P_I)$ where $P_I = I - P_b$ is the projection operator associated with states in which the server is idle. The term $-\theta P_I$ represents a constant flow of probability (with rate θ) from idle states into the absorbing state corresponding to the tagged customer having entered service. The cumulative distribution function of the waiting time is thus given by

$$F(t) = 1 - x_t exp[(Q - \theta P_I)t]e.$$

Since we have already assumed an approximation in which Q is finite dimensional, we can evaluate this expression by applying the randomization method as follows: If λ is the absolute value of the largest diagonal

element of Q, then $P = I + \lambda^{-1}(Q - \theta P_I)$ is substochastic and the series $\{x_t P^k e | k = 0, 1, \ldots\}$ will converge to zero as $k \to \infty$. Then we can expect a reasonable convergence for the Taylor series expansion

$$F(t) = \sum_{k=0}^{\infty} \frac{(\lambda t)^k}{k!} exp(-\lambda t)(1 - x_t P^k e).$$

We can also obtain an approximate upper bound on the error due to truncating the series by applying Stirlings approximation $k! \approx k^k exp(-k)\sqrt{2\pi k}$. If k is larger than λt and large enough to merit the use of Stirling's approximation (about 1% error for $k = 10$ or .1% error for $k = 50$) then

$$\frac{d}{dk} ln\left(\frac{(\lambda t)^k exp(-\lambda t)}{k!}\right) \approx ln\left(\frac{a}{k}\right) - \frac{1}{2k} < 0$$

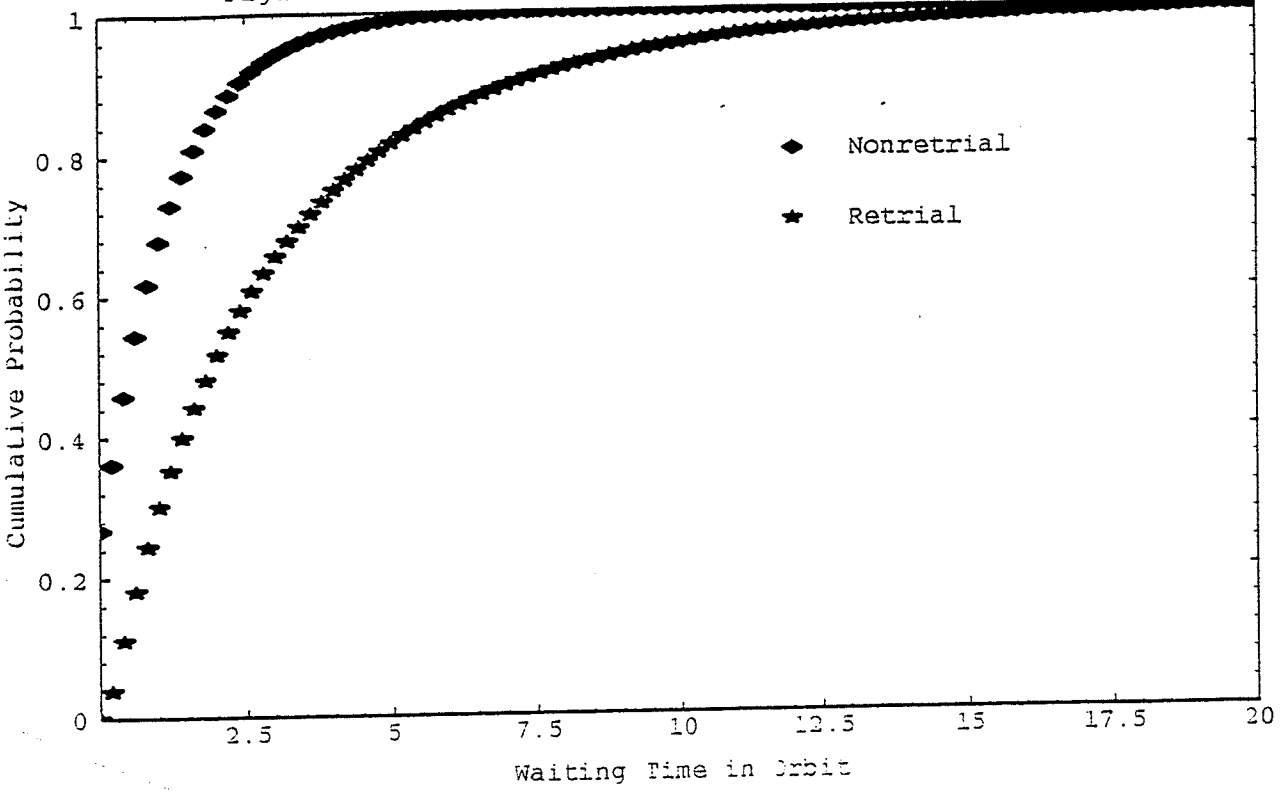
so that the error $\Delta_K F(t)$ due to truncating the series at $k = K$ approximately satisfies

$$\begin{aligned} \Delta_K F(t) &\leq \frac{(\lambda t)^K exp(-\lambda t)}{K!} \sum_{k=K}^{\infty} x_t P^k e \\ &= \frac{(\lambda t)^K exp(-\lambda t)}{K!} x_t P^K (I - P)^{-1} e. \end{aligned}$$

Since this bound is an increasing function of t, we can choose K such that $\Delta_K F(t_{max}) < \epsilon$ where t_{max} is the latest time at which we wish to evaluate $F(t)$ and the approximation will obey the error bound for all smaller times as well.

Figure 1.4 compares the distributions of waiting time for the retrial queue example from section 1.5 and for a nonretrial queue with infinite waiting room and identical arrival and service processes.

Figure 1.4: CDF of Waiting Time (Retrial and Nonretrial)



The problem with this method is that, although it quite straightforward, it requires a long time to implement. In general, the randomization method is not renowned for its speed of implementation. It is normally used for obtaining the transient behavior of queues. If information is required about the transient behavior of a retrial queue during the waiting time of a customer, we can obtain this information without much extra effort. For example, we may be interested in the evolution of measures such as the mean number in the orbit during the waiting time of a customer. The conditional distribution of states at time t into the retrial of a customer given that the customer is still in orbit is given by $yexp[(Q - \theta P_I)t]/(1 - F(t))$ if y is the distribution of states at the arrival time of the customer. Note that we may want to take y different from x_t if we are interested in the waiting time of a customer who arrives when the system is not in equilibrium. For example,

we may be interested in the waiting time as a function of the number of customers in orbit at the time of arrival.

One factor contributing significantly to the computational effort required for this method is the short length of the time step $1/\lambda$. Since λ is the fastest transition rate in the queue, it will almost always be given by $\lambda = N\theta$ the maximum composite retrial rate. We are essentially taking into account every retrial, successful or not, of every customer. However, the very short interretrial times at high levels do not contribute very much to the total retrial time. Thus we can expect to obtain a good approximation by omitting the time that the server is idle when the queue is above a certain level, say N_1. This will clearly yield a lower bound on the waiting time. We can accomplish this by calculating the distribution of waiting time using the generator that describes the evolution of the queue on the subset of states corresponding to idle and busy states below level N_1 and only busy states above levelN_1. We can obtain this generator by applying block Gaussian elimination to eliminate the idle states above level N_1. We accomplish this by making the replacements

$$A_{i0} \leftarrow D_1 \otimes I$$

$$A_{i1} \leftarrow I \otimes S + D_0 \otimes I + (I \otimes S^0)(i\theta I - D_0)^{-1}(D_1 \otimes \beta^T)$$

$$A_{i2} \leftarrow i\theta(i\theta I - D_0)^{-1} \otimes S^0\beta^T$$

for $i \geq N_1$. The time step for calculating the lower bound will be $1/\theta N_1$, which is larger than the original time step. A good rule of thumb for estimating N_1 is that approximately $\sum_{i=N_i}^{N} x_i^0 e$ of the waiting time is discarded (recall x_i^0 is the vector of stationary probabilities corresponding to the idle states at level i, if the queue does not stray very far from equilibrium during the waiting period.

We can obtain an upper bound on the waiting time by considering the queue in which the composite retrial rate for level i is slowed down

to $N_2\theta$ for $i \geq N_2$ and by not allowing the tagged customer to retry above level N_2. Prohibiting retrials for the tagged customer above level N_2 is necessary since, otherwise, slowing down the retrial rate of the other customers would give the tagged customer an advantage and we would not obtain an upper bound. A formal argument establishing this as an upper bound can be obtained from Proposition 1.6 if we modify the partial order (1.17) so that the absorbing state (tagged customer in service) is smaller than all other states and delete the partial order relations between states with different server configurations (idle or busy) below level N_2. The partial order then becomes : $(r, s, t) \leq (r', s', t')$ iff $s = s'$ and $(r < r'$ or $t \leq_s t')$ and $(r' \geq N_2$ or $t, t' = 0$ or $t, t' \neq 0)$. The partial orders between idle and busy states below level N_2 are removed because otherwise, in the mappings $\Theta'_{j\ell}$ defined in (1.20), the transitions from idle states to the lowest state would "cross over "identity transitions in the busy states. This would violate the monotonicity of the $\Theta'_{j\ell}$. We omit the details of the proof.

Both of the above approximations above will reduce the computation time by increasing the step size. We expect that the rate of convergence would be approximately linear in the relative increase in stepsize (N/N_1 or N/N_2) since the mean number of steps in a given time interval is inversely proportional to the size of the step. Although this would represent some time savings, the method is still, by its nature time consuming and faster approximations would be desirable.

A faster approximation can be obtained as follows: The joint distribution of the number of retrials and the inter-retrial times for all of the retrials is given by

$$\begin{aligned} f_k(t_1, \ldots, t_k) =& x_t \left(\theta exp[-\theta t_1] exp[Qt_1] P_b\right) \ldots \\ & \times \left(\theta exp[-\theta t_{k-1}] exp[Qt_{k-1}] P_b\right) \\ & \times \left(\theta exp[-\theta t_k] exp[Qt_k](I - P_b)\right) e. \end{aligned}$$

The joint distribution of the number of retrials and the jth $(j \leq k)$

retrial time is

$$f_k(t_j) = \theta exp[-\theta t_j] x_t (\theta(\theta I - Q)^{-1} P_b)^{j-1} \times exp[Qt_j] P_b (\theta(\theta I - Q)^{-1} P_b)^{k-j-1} (\theta(\theta I - Q)^{-1}(I - P_b))e$$

Because of the presence of the factor $exp[Qt_j]$, it is clear that the conditional distribution of the jth retrial time given that there are k retrials is not exponential. Also, the jth retrial time cannot be considered independent of the number of retrials. However, the effect of $exp[Qt_j]$ is not very important when t_j is large because of the exponential damping factor $\exp[-\theta t_j]$. If the jth retrial time is small, then the evolution of the queue through this relatively short time should not have a great effect on the distribution of the number of retrials which follow it. If we make the zeroth order approximation $exp[Qt_j] \approx I$, we obtain $f_k(t_j) \approx P(n = k)\theta exp[-\theta t_j]$ so that, in our approximation, t_j is exponentially distributed and independent of the number of retrials. If we also assume that all of the retrial times in a sequence ending in absorption (tagged customer enters service) after the kth are independent, we obtain the following approximation for the distribution of waiting time in the orbit:

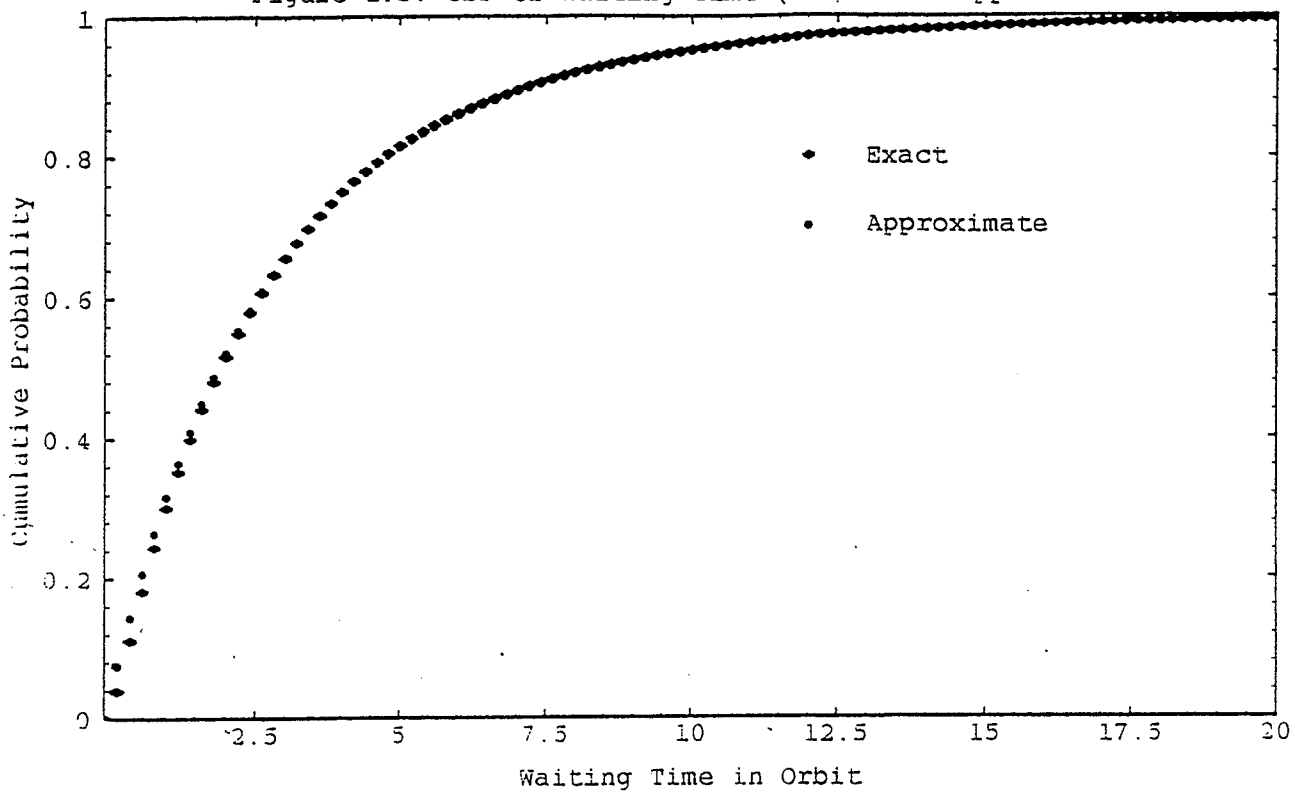
$$p_w(t) \approx P(n = 0)\delta(t) + (1 - P(n = 0)) \sum_{k=1}^{\infty} P(n = k | n > 0) \frac{\theta^k}{(k-1)!} t^{k-1} e^{-\theta t} \quad (1.22)$$

where $\delta(t)$ is the Dirac delta function centered at zero. This assumption of independence should not be too far from the truth since the retrial times are related only by the fact that the server is busy at the end of each of them, except for the last. An exact expression for the conditional waiting time distribution is given by

$$f(t) = \sum_{k=1}^{\infty} x_t \left[\int_0^t dt_1 \int_0^{t-t_1} dt_2 \ldots \int_0^{t-\sum_{j=1}^{k-1} t_j} dt_k \times (\theta exp[(Q - \theta I)t_1] P_b) \ldots (\theta exp[(Q - \theta I)t_{k-1}] P_b) \times exp[(Q - \theta I)t_k](I - P_b)e. \right.$$

Unfortunately, we have no bound on the error for this approximation but it performs remarkably well in the example considered. The exact waiting time distribution is compared to the approximation in figure 1.5.

Figure 1.5: CDF of Waiting Time (Exact and Approximate



1.9 CONDITIONAL WAITING TIME MOMENTS

If only the moments of the waiting time distribution are required, it is possible to obtain these without calculating the entire distribution. The waiting time of a customer not receiving immediate service has a phase type distribution with representation $(x_t, Q-\theta P)I$ and so the jth moment is given by

$$M_j = (-1)^j j! x_t (Q - \theta P_I)^{-j} e.$$

In order to obtain these moments. we need to solve systems of the form $yQ' = \alpha$. This can be done by applying Algorithm 1.3 if we first replace

the recursion for L_i in (1.8) with

$$L_i = -[D_0 - (i+1)\theta I + \gamma_i^T (I \otimes S^0)^{-1}.$$

This requires $O(Nm^3n^2) + O(NMm^2n^2)$ time to calculate the first M moments. Again, $\hat{R}_i$ and R must be reinterpreted as expected sojourn times before the first return to some initial level or the entry into service of our tagged customer.

1.10 LEVEL DEPENDENT EXTENSION: LOCAL AREA NETWORK WITH CSMA PROTOCOL

There are some models of intereste closely related to the MAP/PH/1 retrial queue which posess some level dependence other than that of the composite retrial rate. Notably, models with a finite number of customers, such as models of small computer networks, often posess such level dependence. Coyle and Liu (1985) presented a matrix representation of a network with carrier sense multiple access protocol and collision detection. In this protocol, users check a single shared bus to determine if it is busy before sending a packet onto the bus. If the bus is busy, the user attempts transmission again at some exponentially distributed time later. When a user begins transmission, it requires some time (PH distributed in the Coyle and Liu model) for other users to detect the transmission during which time another user may begin transmitting. This is known as a collision and, when it occurs, both users must retransmit and the bus requires some time (PH distributed) to reset itself and become available for transmission once again. Users which have messages to send retry periodically and are considered busy until the message is sent. We will refer to customers with no message to send as active (since they are presumably active in generating the messages). It is assumed that messages to be sent on the bus arrive to

each non-busy user according to a Poisson process.

We relax the assumption of Poisson arrivals to the active users and allow the time between the appearance of successive messages at each active user to have a two dimensional phase type distribution. The arrival process to the bus then becomes a level dependent MAP. Assuming a two dimensional representation allows us to reduce the size of the representation for the arrival process by lumping together states with the same number of customers in the second phase of their arrival process. The dimension of the arrival process then grows only linearly with the number of active stations. In order to simplify the model, we ignore the possibility of collisions and assume that each user becomes aware of a busy bus immediately. The collisions can be returned to the model without significantly altering the method of analysis, but since our purpose is simply to give an example of a level dependent MAP/PH/1 queue, we omit that aspect from this model.

The Network is modelled as a $PH_2/PH/1//N$ retrial queue. The interarrival time for each of the N sources has a phase type distribution with representation (α, T) where $\alpha = (\alpha_1, \alpha_2)$ and

$$T = \begin{bmatrix} -(t_1 + \gamma_1) & t_1 \\ t_2 & -(t_2 + \gamma_2) \end{bmatrix}.$$

The time required to transmit a message (service time) has a phase type distribution with m-dimensional representation (β^T, S). The retrial rate θ is the reciprocal of the mean time between checks of the bus for a customer in orbit. The state space for the Markov chain which models the network is given by $\{(i, j, k, \ell) | i = 0, 1, \ldots, M; j = 0, 1; k = 0, N - i - \delta_{j1}; \ell = 1, \ldots, m_j\}$ where $m_0 = 1$, $m_0 = m$ and the Kroenecker δ_{j1} is equal to one if $j = 1$ and zero otherwise. The labels i, j, k and ℓ represent respectively the number of customers (stations) in orbit, the number of customers (stations) in service (currently transmitting via the bus), the number of customers in phase two of their two state interarrival process and the phase of service. If we order the state space

so that the labels (i, j, k, ℓ) appear in lexicographic order, the generator for the Markov chain has the following structure:

$$Q = \begin{bmatrix} A_{0,1} & A_{00} & 0 & \cdots & \\ A_{12} & A_{11} & A_{10} & 0 & \cdots \\ 0 & \ddots & \ddots & \ddots & \\ \vdots & & \ddots & \ddots & A_{N-2,0} \\ & & & A_{N-1,2} & A_{N-1,1} \end{bmatrix}$$

where:

$$A_{i0} = \begin{bmatrix} 0_{N-i+1 \times N-i} & 0 \\ 0 & D_{N-i-1} \otimes I \end{bmatrix}$$

$$A_{i2} = \begin{bmatrix} 0_{N-i+1 \times N-i+2} & i\theta I_{N-i+1} \otimes \beta^T \\ 0 & 0_{m(N-i) \times m(N-i+1)} \end{bmatrix}$$

$$A_{i1} = \begin{bmatrix} C_{N-i} - i\theta I_{N-i+1} & D_{N-i} \otimes \beta^T \\ U_{N-i} \otimes S^0 & I_{N-i} \otimes S + C_{N-i-1} \otimes I \end{bmatrix} \tag{1.23}$$

$$A_{N-1,1} = \begin{bmatrix} C_1 - i\theta I_2 & D_1 \otimes \beta^T \\ U_1 \otimes S_0 & S \end{bmatrix}$$

and $S^0 = -Se$. This is the generator for a MAP/PH/1//N retrial queue with level dependent arrival process represented by (C_i, D_i) when there are i active customers. Note that if there are $N-1$ customers in orbit and one active customer, the active customer goes directly into service if he attempts to seize the bus before the next retrial from orbit. If there are $N-1$ customers in orbit and one in service, then the customer in service becomes active when he completes his service. There is thus no way for all N customers to be in orbit at the same time. This is why there is no level N represented in the generator.

The matrices C_i, D_i and U_i are given by

$$C_i = \begin{bmatrix} c_{i01} & c_{i00} & & \\ c_{i12} & \ddots & \ddots & \\ & \ddots & \ddots & c_{i,i-1,0} \\ & & c_{ii2} & c_{ii1} \end{bmatrix} \in \Re_{(i+1)\times(i+1)}$$

$$D_i = \begin{bmatrix} d_{i01} & & & \\ d_{i12} & \ddots & & \\ & \ddots & d_{i,i-1,1} & \\ & & d_{ii2} & \end{bmatrix} \in \Re_{(i+1)\times i}$$

$$U_i = \begin{bmatrix} \alpha_1 & \alpha_2 & & \\ & \ddots & \ddots & \\ & & \alpha_1 & \alpha_2 \end{bmatrix} \in \Re_{(i+1)\times(i+2)}$$

where I_i is the i-dimensional identity matrix and the scalars c_{ijk} and d_{ijk} are defined by

$$c_{ij0} = (i-j)t_1 \qquad c_{ij2} = jt_2 \qquad c_{ij1} = -(j(t_2+\gamma_2)+(i-j)(t_1+\gamma_1))$$

$$d_{ij2} = j\gamma_2 \qquad d_{ij1} = (i-j)\gamma_1.$$

If there are i active customers and j of them are in phase 2 of their arrival process, then $i-j$ are in phase 1. c_{ij0} is the rate of phase 1 customers switching to phase 2 and c_{ij2} is the rate of phase 2 customers switching to phase 1. d_{ij1} and d_{ij2} are the arrival rates of customers in phase 1 and phase 2 respectively. D_i has no superdiagonal elements because the number of customers in phase 2 of their arrival process cannot increase due to a customer arriving to join the orbit or service. This number will remain the same (d_{ij1}) if a phase 1 customer joins the orbit or service and will decrease by one (d_{ij2}) if a phase 2 customer joins the orbit or service. The form of U_i arises from the fact that the number of phase 2 customers will remain the same if a customer completes service and begins an arrival process in phase 1 (with probability α_1) and will increase by one if a customer completes service and joins the active customers in phase 2 (with probability α_2) of their arrival processes.

The solution method is similar to the one used in the previous sections except that the recusion relations (1.7) and (1.8) are altered, to take into account the form of the level dependent arrival process.

After substituting from (1.23), we obtain the following recursion for the M_i and δ_i^T:

$$\delta_0 = 0$$

$$M_0 = 0$$

$$\delta_{i+1}^T = (i+1)\theta[\nu_i^T + J_i L_i \gamma_i^T]$$

$$M_{i+1} = (i+1)\theta J_i L_i \tag{1.24}$$

where

$$\nu_i^T = -(I_{N-i} \otimes \beta^T)(I_{N-i} \otimes S + C_{N-i-1} \otimes I)^{-1}$$

$$J_i = \nu_i^T (U_{N-i} \otimes S^0)$$

$$\gamma_i^T = -[D_{N-i} \otimes \beta^T + \delta_i^T (D_{N-i} \otimes I)](I_{N-i} \otimes S + C_{N-i-1} \otimes I)^{-1}$$

$$L_i = -[C_{N-i} - i\theta I_{N-i+1} + \gamma_i^T (U_{N-i} \otimes S^0)]^{-1}. \tag{1.25}$$

Since the matrices $I_{N-i} \otimes S + C_{N-i-1} \otimes I$ are block tridiagonal, we can use block gaussian elimination to apply its inverse to a row vector of the appropriate dimension in $O((N-i)m^3) + O((N-i)m^2)$ time instead of the $O((N-i)^3 m^3)$ time required to invert a general matrix of the same dimension. The ith step of the algorithm thus requires $O((N-i)^3) + O((N-i)m^3) + O((N-i)^2 m^2)$ time and the entire algorithm requires $O(N^4) + O(N^2 m^3) + O(N^3 m^2)$ time. Since we only have to invert matrices of dimension N or m instead of matrices of dimension Nm, software limitations on the size of matrices which can be inverted may be less of a problem.

In order to calculate the conditional probabilities $P(n = k | n > 0)$ of the number of retrials we must be able to solve equations of the form $y(\theta I - Q) = \alpha$. This can be done by making the substitutions

$C_i \leftarrow C_i - \theta I_{i+1}$ and then applying Algorithm 1.3 with the following substitution for the matrices X_i and Y_i:

$$Y_i = (-A'_{i1})^{-1} A_0 = \begin{bmatrix} 0 & L_i \gamma_i^T (D_{N-i-1} \otimes I) \\ 0 & T_i [I_{m(N-i)} + (U_{N-i} \otimes S^0) L_i \gamma_i^T] \end{bmatrix}$$

$$X_i = (-A'_{i1})^{-1}$$

$$= \begin{bmatrix} L_i & L_i \gamma_i^T \\ T_i (U_{N-i} \otimes S^0) L_i & T_i [I_{m(N-i)} + (U_{N-i} \otimes S^0) L_i \gamma_i^T \end{bmatrix}$$

$$T_i = -(I_{N-i} \otimes S + C_{N-i-1} \otimes I)^{-1} \qquad i = 0, \ldots, N-1$$

and $T_{N-1} = -S^{-1}$ We must apply the algorithm M times to calculate $P(n = k | n > 0)$ for $k = 1, 2, \ldots, M$. Since the first part of Algorithm 1.3, which has a time complexity of $O(N^4) + O(N^2 m^3) + O(N^3 m^2)$, only needs to be executed once, the time required to get all M probabilities is $O(N^4) + O(N^2 m^3) + O(N^3 m^2) + O(MN^2 m^2) + O(MN^3 m)$.

In order to calculate the moments of the waiting time distribution, we must replace the recusion for L_i in (1.25) with

$$L_i = -[C_{N-i} - (i+1)\theta I_{N-i+1} + \gamma_i^T (U_{N-i} \otimes S^0)]^{-1}$$

and apply algorithm 1.4 to solve systems of the form $y(Q - \theta P_I) = \alpha$.

We do not include any further detailed analysis of the model since our purpose was merely to give an example of modifications required to adapt the method to queues with level dependence. This provides an idea of the range of models to which the method is applicable. If the number of stations is large (more than 20-30 or so), an adaptation of the methods in Chapter Four may provide a good approximation which requires significantly less computational effort than the exact solution . We leave the description of this approximation method to Chapter Four.

1.11 CONCLUSION

We have derived a numerical method for obtaining the stationary distribution of states and the distributions of waiting time and the number of retrials for a MAP/PH/1 retrial queue. We have also obtained a relatively fast method for obtaining the moments of the waiting time. This extends previous models by allowing the inclusion of correlation in the arrival process as well as nonexponentially distributed interarrival times.

We have also derived a method for obtaining bounds on probability lost due truncation by considering approximations which are monotonic and homogeneous above some level. In doing so, we have identified a class of phase type distributions which is closed under convolution and which, we believe , may give rise to monotone processes when it appears in models other than the MAP/PH/1 retrial queue as well. The strategy of using the M/G/1 paradigm and focusing on the duals $\hat{R}_i$ of the rate matrices R_i of the GI/M/1 paradigm allows us to delay decisions about where to truncate or approximate level dependent QBD processes until we have sufficient information to guarantee a certain level of accuracy in the approximation. This strategy should prove useful in the analysis of general level dependent QBD processes. We have also demonstrated that the method of this chapter can be extended to QBD processes with level dependence other than the composite retrial rate, as in the case of the CSMA LAN model. In the following chapter, we present some analytical results for the special case of the M/PH/1 retrial queue which are not available in the more general model. In subsequent chapters, we show how the methods of this chapter can be adapted and extended to other related retrial models such as retrial queues with buffers and retrial queues with non-exponential retrial times.

CHAPTER TWO

THE M/PH/1 RETRIAL QUEUE

1.1 INTRODUCTION

Keilson et al. (1968) derived the generating function of the stationary distribution of queue length for the M/G/1 retrial queue with no loss or buffer. Obtaining the distribution itself, however, can be cumbersome in practice. The expression for the generating function involves an integral which may be difficult to solve explicitly and the z-transform thus obtained may also be difficult to invert.

Greenberg (1989) developed a method to approximate the steady state distribution by assuming that returning customers see time averages. Greenberg and Wolff (1987) have shown that this assumption leads to an upper bound on server utilization.

De Kok (1984) derived separate expressions for the generating functions of the number of customers in the queue when the server is busy and when the server is idle. Artalejo (1993) solved the integral in the latter expression explicitly for the special case of the M/H_2/1 retrial queue, inverted the transform and used the Kolmogrov equations directly to solve for the busy probabilities in terms of the idle ones.

De Kok also developed a numerical method for some special cases including service time distributions which are finite mixtures of Erlangs with common intensity. Since any phase type distribution can be considered an infinite mixture of this form, more general phase type distributions could be approximated by truncating the mixture series. In fact, any distribution on $(0, \infty)$ can be approximated arbitrarily closely by a distribution of the form considered by de Kok, however the dimension of the representation may become impractically large. Since

current methods (see Asmussen and Nerman (1991), Bobbio et al (1980) and Johnson and Taaffe (1988))for fitting phase type distributions use either general phase type distributions or other special cases (Coxian, mixture of Erlangs with different intensities, triangular representations) it would be of more practical use to have a simple method applicable to retrial queues with general phase type service time distributions.

Neuts and Ramalhoto (1990) consider a service model in which the server is required to search for customers. Though the context is different from that of the retrial queue, the model is equivalent to an M/G/1 retrial queue where all arriving customers go immediately into orbit even when the server is idle. The generating function is derived and a numerical method of obtaining the stationary distribution is given. This method (as Neuts and Ramalhoto point out) can be adapted to obtain the distribution for the M/G/1 retrial queue and does not require numerical integration if the service time distribution is of phase type.

Although an analytically explicit solution is presented in Neuts and Ramalhoto (1984), the method suggested is more cumbersome than necessary for distributions of phase type because they arrive at the stationary distribution via the generating function whereas the balance equations are more easily solved directly. The suggested method requires calculation of the distribution $\{a_\nu, \nu \geq 0\}$ of the number of arrivals during a service as well as the convolution of an infinite number of Poisson distributed variables and a number of other convolutions. The effort required to calculate the stationary distribution directly is similar to the first of these steps and no convolutions are necessary. A more direct method is applied in Neuts and Rao (1990) to investigate the M/M/s retrial queue. Although generating functions are avoided, the method differs somewhat from ours in that an iterative scheme is applied to solve the balance equations whereas we apply a direct method.

If we model the M/PH/1 retrial queue as a discrete space Markov

process, we find that the generator is block tridiagonal with subdiagonal blocks of rank one. Generators of this form admit an explicit matrix product form solution for the stationary distribution, when it exists. Though a stability condition has already been established for the more general M/G/1 retrial queue, for the sake of completeness we show how this stability condition determines the convergence or divergence of the matrix series which arises in the normalization.

Falin (1991) obtained a closed form expression for the Laplace transform of the waiting time distribution and for the generating function of the number of retrials per customer in the M/G/1 queue. However, it may be difficult in practice to solve the integrals and invert the transforms which appear in these expressions. If the service time distribution is of phase type, it is possible to obtain the distribution of the number of retrials numerically using only matrix multiplication and a single matrix inversion. The waiting time distribution and moments can be obtained by applying the methods outlined in Chapter One.

2.2 STATIONARY DISTRIBUTION

Consider a retrial queue with arrival rate λ, retrial rate θ per customer and service time distribution represented by (β, S). The state space is the set

$$E = \{-1\} \cup \{(0,j)|j = 1,\ldots,m\} \cup \{(i,j)|i = 1,2,\ldots; j = 0,1,\ldots,m\}.$$

The indices i and j represent respectively the number of customers in orbit and the phase of service ($j = 0$ if the server is idle) and $\{-1\}$ represents the state of the system when the orbit is empty and the server idle. m is the dimension of the service time distribution.

The generator of the corresponding Markov process is of the form

$$Q = \begin{bmatrix} A_{-1,1} & A_{-1,0} & 0 & 0 & 0 & \dots \\ A_{02} & A_{01} & A_{00} & 0 & 0 & \dots \\ 0 & A_{12} & A_{11} & A_0 & 0 & \dots \\ 0 & 0 & A_{22} & A_{21} & A_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{2.1}$$

where:

$$A_{-1,1} = -\lambda \qquad A_{-1,0} = \lambda\beta^T$$

$$A_{02} = S^0 \qquad A_{01} = S - \lambda I \qquad A_{00} = [0 \quad \lambda I]$$

$$A_{12} = \begin{bmatrix} \theta\beta^T \\ 0 \end{bmatrix} \qquad A_0 = \lambda \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} \tag{2.2}$$

and for $i \geq 1$:

$$A_{i2} = \begin{bmatrix} 0 & i\theta\beta^T \\ 0 & 0 \end{bmatrix} \qquad A_{i1} = \begin{bmatrix} -(i\theta + \lambda) & \lambda\beta^T \\ S^0 & S - \lambda I \end{bmatrix}. \tag{2.3}$$

For any vector $x \in \Re^\infty$ we partition x according to $x = [x_{-1}, x_0, x_1, \dots]$ where $x_{-1} \in \Re$, $x_0 \in \Re^m$ and $x_i \in \Re^{m+1}$ for $i \geq 1$. We also define $x_{(i,j)} \in \Re$ to be the jth (scalar) component of the vector x_i. The generator Q is block tridiagonal with subdiagonal blocks of rank one. Ramaswami and Latouche (1986) gave the stationary distribution for the special case of generators of this type for which the blocks along a diagonal are also identical, except for the first. Snyder and Stewart (1985) gave the solution for generators with rank one subdiagonal blocks and superdiagonal blocks which are diagonal. The method used in both these cases is similar to that used in Neuts (1981) to solve the M/PH/1 queue. The following two propositions summarize the extension of this method to the general case of block tridiagonal generators with rank one subdiagonal blocks. The result is a particular case of a general result for the matrix multiplicative form of stationary vectors for generators with block Hessenberg structure and rank

one subdiagonal blocks (see Basharin and Naumov (1983) and Naumov (1985)). It is the GI/M/1 analog of Proposition 1.3.

Proposition 2.1: *If the generator for a positive recurrent Markov process has the form*

$$Q = \begin{bmatrix} A_{01} & A_{00} & 0 & 0 & 0 & \dots \\ A_{12} & A_{11} & A_{10} & 0 & 0 & \dots \\ A_{23} & A_{22} & A_{21} & A_{20} & 0 & \dots \\ A_{34} & A_{33} & A_{32} & A_{31} & A_{30} & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{2.4}$$

where each A_{ij} is an $m_i \times m_{i-j+1}$ matrix, the stationary distribution x satisfies $x_{i+1} = x_i \quad R_i$ where the $m_i \times m_{i-j+1}$ matrices R_i satisfy:

(i) $0 = \sum_{k=0}^{\infty} \quad R_i^{(k)} A_{i+k,k}$

(ii) $A_{i0}e = \sum_{k=1}^{\infty} \quad R_i^{(k)} \sum_{\nu=k+1}^{k+1+i} \quad A_{i+k,\nu}e$

with $R_i^{(k)} = \prod_{j=i}^{i+k-1} R_i$. The component $[R_i]_{jk}$ is the expected time spent in the state $(i+1,k)$ before the first return to level i measured in units of duration equal to the mean sojourn time in the state (i,j).

Proof: The proof is analogous to that of Proposition 1.3. □

The stationary distribution is thus given by

$$x_j / x_{-1} = \prod_{i=-1}^{j-1} R_i \tag{2.5}$$

where the matrices R_i are determined by the following:

Proposition 2.2: *If Q as defined in (2.4) is block tridiagonal and satisfies $A_{i2} = \gamma_i \alpha_{i-1}^T \qquad i = 1, 2, 3, \ldots$ where γ_i and α_{i-1} are non-zero column vectors of the appropriate length satisfying $\gamma_i, \alpha_i \geq 0$ and $\alpha_i^T e = 1$, then, for $i \geq 0$:*

$$R_i = -A_{i0}(A_{i+1,1} + A_{i+1,0} e \alpha_{i+1}^T)^{-1}.$$

Proof: For block tridiagonal generators, part (i) of Proposition 2.1 reduces to

$$0 = A_{i0} + R_i \, A_{i+1,1} + R_{i \; i+1}R \, A_{i+2,2} \tag{2.6}$$

and part (ii) reduces to

$$A_{i0} e = R_i A_{i+1,2} e. \tag{2.7}$$

If we replace i in (2.7) with $i+1$ and postmultiply by α_{i+1}^T, we obtain

$$A_{i+1,0} e \alpha_{i+1}^T = {}_{i+1}R A_{i+2,2}.$$

Substituting this into (2.6) yields

$$R_i(A_{i+1,1} + A_{i+1,0} e \alpha_{i+1}^T) = -A_{i0}.$$

Let $B_i = (A_{i1} + A_{i0} e \alpha_i^T)$ and let $[B_i]_{jk}$ denote the (j,k) element of B_i. Then B_i satisfies

(a) $[B_i]_{jk} \geq 0$ if $j \neq k$.
(b) $B_i e = \gamma_i$ (≤ 0 and $\neq 0$).

(a) and (b) are sufficient to conclude that B_i is a stable matrix (see Marcus and Minc (1964) p. 158). Thus all the eigenvalues of B_i are strictly negative and the inverse B_i^{-1} exists so that we can write

$$R_i = -A_{i,0}(A_{i+1,1} + A_{i+1,0} e \alpha_{i+1}^T)^{-1}. \square$$

Substituting the coefficient matrices in (2.2) and (2.3) into the expression for R_i in Proposition 2 yields $R_{-1} = \beta^T R$, $R_0 = [\frac{\lambda}{\theta}e, R + \frac{\lambda}{\theta}e\beta^T R]$ and :

$$R_i = \begin{bmatrix} 0 & 0 \\ \frac{\lambda}{(i+1)\theta}e & R + \frac{\lambda}{(i+1)\theta}e\beta^T R \end{bmatrix} \qquad i = 1, 2, \ldots \tag{2.8}$$

where $R = -\lambda(S - \lambda I + \lambda e\beta^T)^{-1}$ is the rate matrix derived in Neuts (1981) for the M/PH/1 queue with an infinite number of waiting positions. The first component of the relation $x_{i+1} = x_i R_i$ corresponds to equation (2.6) in De Kok (1984). Note that, in the limit as θ approaches infinity, the probability $\sum_{i=1}^{\infty} x_{(i,0)}$ that the server is idle with customers in the queue approaches zero and the probabilities $x_{(i,j)}$ with $j \neq 0$ approach their counterparts from the M/PH/1 queue with an infinite number of waiting positions. This is, of course, expected since, if the expected time $1/\theta$ which a customer waits before he retries is zero, the queue reduces to an M/PH/1 queue with random service order.

2.3 POSITIVITY AND STABILITY

If the representation (β, S) is irreducible, the matrix $(S - \lambda I + \lambda e\beta^T)$ is stable and irreducible so that its inverse (and thus R) is positive. This guarantees the positivity of x. The stability condition $\rho = \lambda/\mu < 1$, where $\mu^{-1} = -\beta S^{-1}e$ is the expected service time, is well known (see Falin (1990)). We shall prove that this guarantees the convergence of the matrix series obtained by summing the right hand side of (2.5). The method we use is an extension of Theorem 3.2.2 in Neuts (1981).

We define the quantities r_i $(i = 1, 2, \ldots)$ by

$$r_i = -\alpha_{i+1}^T (A_0 + A_{i+1,1})^{-1} A_0 e.$$

We shall show that if the limit of the sequence $\{r_i\}$ is less than 1 then xe is finite so that the probability vector x is normalizable. The following

proposition provides an alternative expression for R_i which we require to formulate a bound on the spectral radius $sp(R_i)$.

Proposition 2.3: If the generator for a Markov Process satisfies the conditions of Proposition 2.2, then the corresponding rate matrices R_i defined in Proposition 2.1 satisfy:

$$R_i = -A_0 A_{i+1,1}^{-1} + \frac{A_0 A_{i+1,1}^{-1} A_0 e \alpha_i^T A_{i+1,1}^{-1}}{\alpha_i^T A_{i+1,1}^{-1} \gamma_i}.$$

Proof:

$$\begin{aligned}
(A_{i+1,1} + A_0 e\alpha_i^T)^{-1} &= A_{i+1,1}^{-1}[I + A_0 e\alpha_i^T A_{i+1,1}^{-1}]^{-1} \\
&= A_{i+1,1}^{-1}\left[I + \sum_{\nu=1}^{\infty}(-A_0 e\alpha_i^T A_{i+1,1}^{-1})^{\nu}\right] \\
&= A_{i+1,1}^{-1}\left[I - A_0 e\alpha_i^T A_{i+1,1}^{-1}\sum_{\nu=0}^{\infty}(-\alpha_i^T A_{i+1,1}^{-1} A_0 e)^{\nu}\right] \\
&= A_{i+1,1}^{-1}\left[I - \frac{A_0 e\alpha_i^T A_{i+1,1}^{-1}}{1 + \alpha_i^T A_{i+1,1}^{-1} A_0 e}\right]
\end{aligned}$$

Now we have

$$\begin{aligned}
1 = \alpha_i^T e = \alpha_i^T A_{i+1,1}^{-1} A_{i+1,1} e &= -\alpha_i^T A_{i+1,1}^{-1}(A_0 e + \gamma_i) \\
&= -\alpha_i^T A_{i+1,1}^{-1} A_0 e - \alpha_i^T A_{i+,11}^{-1}\gamma_i.
\end{aligned}$$

Thus

$$\begin{aligned}
R_i &= -A_0(A_{i+1,1} + A_0 e\alpha_i^T)^{-1} \\
&= -A_0 A_{i+1,1}^{-1} - \frac{A_0 A_{i+1,1}^{-1} A_0 e\alpha_i^T A_{i+1,1}^{-1}}{\alpha_i^T A_{i+1,1}^{-1}\gamma_i}.
\end{aligned}$$

□

Corollary 2.3: *If* $A_{i+1,1} + A_0 e\alpha_i^T$ *is irreducible and* $A_0 \neq 0$ *then* $sp(R_i) > sp(-A_0 A_{i+1,1}^{-1})$.

Proof: Let $(A_{i+1,1} + A_0 e\alpha_i^T)$ be irreducible. Since it is also a stable matrix, its inverse is negative and thus R_i is positive. Clearly $R_i \geq -A_0 A_{i+1,1}^{-1}$ since the second term for R_i in Proposition 3 is non-negative. Suppose that $R_i = -A_0 A_{i+1,1}^{-1}$. Then $A_0 A_{i+1,1}^{-1}$ is negative and $A_0 A_{i+1,1}^{-1} A_{i0} e\alpha_i^T A_{i+1,1}^{-1} = 0$. But this is impossible since $A_0 A_{i+1,1}^{-1} A_0 e$ must be negative if $A_0 A_{i+1,1}^{-1}$ is negative and $A_0 e \neq 0$. Thus we have $R_i \geq -A_0 A_{i+1,1}^{-1}$ and $[R_i]_{jk} \neq [-A_0 A_{i+1,1}^{-1}]_{jk}$ for some element (j,k). This is sufficient to conclude that $sp(R_i) > sp(-A_0 A_{i1}^{-1})$. □

We can now relate r_i to $sp(R_i)$.

Proposition 2.4: Let $r_i = -\alpha_{i+1}^T (A_0 + A_{i+1,1})^{-1} A_0 e$ where α_{i+1}, A_0 and $A_{i+1,1}$ are components of a generator satisfying the conditions of Proposition 2.2. If $r_i < 1$ then $sp(R_i) < 1$. If $r_i > 1$ then $sp(R_i) > 1$.

Proof: Let u be the left eigenvector of R_i corresponding to $\eta = sp(R_i)$. Then the equation $u\, R_i = \eta u$ leads to

$$u = -\eta (u A_0 e) \alpha_{i+1}^T [A_0 + \eta A_{i+1,1}]^{-1}.$$

Postmultiplying by $A_0 e$ yields

$$\begin{aligned} 1 &= -\eta \alpha_{i+1}^T [A_0 + \eta A_{i+1,1}]^{-1} A_0 e \\ &= \eta \alpha_{i+1}^T [(-A_0 A_{i+1,1}^{-1}) - \eta I]^{-1} A_{i+1,1}^{-1} A_0 e\,. \\ &= \Phi(\eta) \end{aligned}$$

where $\Phi(s) = s\alpha_{i+1}^T[A_o + sA_{i+1,1}]^{-1}A_0e$ is defined on the domain $D(\Phi) = \left(sp(-A_0A_{i+1,1}^{-1}), \infty\right]$. For $s \in D(\Phi)$ the matrix $G(s) = [(-A_0A_{i+1,1}^{-1}) - sI]$ is stable so that, by Theorem 1 on page 250 of Bellman (1970),

$$lim_{t\to\infty}\ exp[G(s)t] = 0$$

and so we can write

$$\Phi(s) = -\int_0^\infty \eta\alpha_{i+1}^T\ \boldsymbol{exp}[\boldsymbol{G}(s)t]A_{i+1,1}^{-1}A_0edt$$
$$= \int_0^\infty s\ exp(-st)f(t)dt$$

where $f(t) = -\alpha_{i+1}^T\ exp(-A_0A_{i+1,1}^{-1}t)A_{i+1,1}^{-1}A_0e$. Note that

$$f'(t) = \alpha_{i+1}^TA_0A_{i+1,1}^{-2}A_0\ exp(-A_0A_{i+1,1}^{-1}t)e \geq 0$$

since $A_{i+1,1}^{-1} \leq 0$. Thus $\Phi(s)$ is non-increasing. Let u' be a left eigenvector of $-A_0A_{i+1,1}^{-1}$ with eigenvalue ν. Then we must have

$$u'[A_0 + \nu A_{i+1,1}^{-1}] = 0$$

but the matrix $A_0 + \nu A_{i+1,1}$ is stable (and thus nonsingular) if $\nu \geq 1$ which implies that $sp\left(-A_0A_{i+1,1}^{-1}\right) < 1$ and $1 \in D(\Phi)$. If $r_i = \Phi(1) < 1$ and $\Phi(\eta) = 1$, we must have $\eta < 1$ since Φ is nonincreasing. Similarily, if $r_i > 1$, we must have $\eta > 1$. □

The following lemma and proposition establish conditions for the invariant probability vector x to be normalizable.

Lemma 2.5: *Let $\{R_i : i = 1, 2, \ldots\}$ be a sequence of $n \times n$ complex matrices. If there exists an integer M and a real number $\xi < 1$ (> 1) such that $sp(R_i) \leq \xi$ $(\geq \xi)$ for all $i \geq M$ then the series*

$$S = \sum_{j=1}^{\infty}\prod_{i=1}^{j} R_i$$

is convergent (divergent).

Proof: See Section 1.7.

Proposition 2.5: *If the limit* $\rho = \lim_{i\to\infty} r_i$ *exists, then the Markov chain with generator* Q *is positive recurrent if* $\rho < 1$ *and non-reccurent if* $\rho > 1$

Proof: The proposition follows directly from Proposition 4 and Lemma 2.5.

For the M/PH/1 retrial queue, we have $r_i = \frac{\lambda}{\mu}\left(1 + \frac{1}{(i+1)\theta}\right)$,where $\mu = -\beta^T S^{-1} e$ is the expected service time, so that $\rho = \frac{\lambda}{\mu}$.

2.4 EXTENSION #1: M/M/1 RETRIAL QUEUE WITH GEOMETRIC LOSS

Consider an M/M/1 retrial queue in which customers leave the system with probability r each time they retry for service and fail. Falin (1980) obtained the partial generating function for the joint distribution of the number in service and the number in orbit. Neuts and Ramalhoto (1984) considered a similar model where the server is required to search for customers and where customers leave the pool at a rate proportional to the number of customers present. They obtained the partial generating functions for the number in the orbit when the server is busy and when the server is idle.

The reason that this model is tractable only for the case of exponential service times is rather transparent from the point of view of the

matrix analytical method. The subdiagonal blocks of the generator associated with the M/PH/1 retrial queue with geometric loss are given by

$$A_{i2} = \begin{bmatrix} 0 & i\theta\beta^T \\ 0 & i\theta r I \end{bmatrix} \qquad i = 2, 3, \ldots,$$

where (β, S) is a representation of the service time distribution. The rank of the subdiagonal blocks of the generator are equal to the dimension of this representation so that, if the service times are exponentially distributed, the subdiagonal blocks have rank one. The method of Section 2.1 can then be used to obtain explicit results for the stationary distribution, thus avoiding the inversion of the generating functions.

The generator associated with this model is given by Q in (2.1) with:

$$A_{-1,1} = -\lambda \qquad A_{-1,0} = \lambda$$

$$A_{02} = \mu \qquad A_{01} = -(\mu + \lambda) \qquad A_{00} = [0 \quad \lambda]$$

$$A_{12} = \begin{bmatrix} \theta \\ 0 \end{bmatrix} \qquad A_0 = \begin{bmatrix} 0 & 0 \\ 0 & \lambda(1-r) \end{bmatrix}$$

and, for $i \geq 1$:

$$A_{i1} = \begin{bmatrix} -(\lambda + i\theta) & \lambda \\ \mu & -(\mu + i\theta r + \lambda(1-r)) \end{bmatrix},$$

where μ is the service rate. Substituting into the expression above for r_i yields

$$r_i = \frac{\lambda(1-r)}{(i+1)\theta\left[r + \frac{\mu}{\lambda+(i+1)\theta}\right]}.$$

Since $\rho = \lim_{i\to\infty} r_i = 0$, the queue is always stable as noted by Falin [10]. The rate matrices are given by $R_{-1} = \lambda/\mu$ and

$$R_0 = \frac{r_0}{1-r}\left[\frac{\mu}{\lambda+\theta}, 1\right]$$

$$R_i = r_i \begin{bmatrix} 0 & 0 \\ \frac{\mu}{\lambda+(i+1)\theta} & 1 \end{bmatrix} \qquad ; i \geq 1.$$

The stationary vector x is given by $x/x_{-1} = [1, y'_0, x'_1, y'_1, x'_2, y'_2, \ldots]$, where

$$
\begin{aligned}
y'_0 &= \lambda/\mu \\
y'_1 &= \frac{\lambda}{\mu}\frac{r_0}{1-r} \\
y'_i &= \frac{\lambda}{\mu}\frac{r_0}{1-r}\prod_{j=1}^{i-1} r_j \qquad ; i \geq 1 \\
x'_i &= \frac{\mu r_i}{\lambda + i\theta} y'_{i-1}.
\end{aligned}
$$

2.5 EXTENSION #2: LEVEL DEPENDENCE

The result of Proposition 2.1 depends only on the tridiagonal structure of the generator Q and on the rank of the subdiagonal blocks. Our result for the stationary distribution can thus be extended to include arbitrary state dependence of the arrival and retrial rates and of the service time distribution. If these are given by λ_i, θ_i and (β_i, S_i) respectively when there are i customers in orbit, we need only add the subscript $i+1$ to λ, θ and (β, S) where they appear in (2.8) (including the expression for R) and multiply the right hand side of (2.8) by λ_i/λ_{i+1}.

The results concerning the stability condition, however, also depend on the fact that the superdiagonal blocks are identical except for the first. This property was required to show that $1 \in D(\Phi)$ in Proposition 2.4. Other methods may be required to take into account state dependent arrival rates. It is a simple matter, however, to extend these results to the case in which the retrial rates and service time distributions are state dependent but the arrival rates are not. In this case we have:

$$
r_i = \frac{\lambda}{\mu_{i+1}}\left(1 + \frac{1}{(i+1)\theta_{i+1}}\right).
$$

There are three state dependent models which we feel may merit some future consideration. The first concerns computer communications networks. If the network is small, it may be necessary to model the system as one with a finite number of customers. In this case we have $\lambda_i = \lambda_0(1 - 1/N)$ where N is the number of customers. Though the arrival rate is state dependent, there is no problem concerning stability conditions because the state space is finite.

The second model includes state dependent service time distributions. Consider a model in which the server must do some work to process requests for service which arrive when he is busy. In this case, we would expect the service times to be longer when there are a large number of customers in the orbit since the server must spend time sending busy signals to these customers. Of course, some mechanism must be available to increase the service rate when the number of customers in the orbit grows too large so that the system can reach stability.

The third model includes state dependent retrial rates. This type of behavior may occur if the system has limitations on the number of retrials which it can process simultaneously. This may occur with telephone traffic, for example, where the time to complete an unsuccessful call and receive a busy signal may increase when the system is overloaded.

2.6 CONCLUSION

We have obtained analytical expressions for the rate matrices associated with the M/PH/1 retrial queue and provided sufficient condition for the convergence of the matrix series which appears in the normalization condition for the stationary distribution of states. The more general numerical methods of Chapter One can be applied to obtain the distributions of the waiting time and number of retrials. We also provided an explicit solution for the stationary distribution of the M/M/1

retrial queue with geometric loss which provides some insight into why the problem is tractable for exponential service times. We have noted that the method used is well suited to dealing with some models which include state dependent parameters.

2.7 PROOF OF LEMMA 2.5

Lemma 2.5: *Let* $\{R_i : i = 1, 2, \ldots\}$ *be a sequence of* $n \times n$ *complex matrices. If there exists an integer* M *and a real number* $\xi < 1$ (> 1) *such that* $sp(R_i) \leq \xi$ $(\geq \xi)$ *for all* $i \geq M$ *then the series*

$$S = \sum_{j=1}^{\infty} \prod_{i=1}^{j} R_i$$

is convergent (divergent).

Proof:

(i) convergence

For each matrix R_i let S_i be the $n \times n$ matrix which reduces (via a similarity transformation) R_i to its Jordan canonical form $\tilde{R}_i$ (see Bellman 1970). Then

$$_i\tilde{R} = S_i \, R_i \, S_i^{-1} = \begin{bmatrix} \lambda_1 & a_1 & 0 & 0 & \ldots \\ 0 & \lambda_2 & a_2 & 0 & \ldots \\ 0 & 0 & \ddots & \vdots & \ldots \\ 0 & \ldots & 0 & \lambda_{n-1} & a_{n-1} \\ 0 & 0 & \ldots & 0 & \lambda_n \end{bmatrix}$$

where $a_j \in \{0, 1\}$ and λ_i $(i = 1, 2, \ldots, n)$ is an eigenvalue of R_i. For all $y \in C_n$ we define the norm

$$\|y\| = \max_{1 \leq j \leq n} |y_j|$$

and the associated operator norm

$$\|A\| = \sup_{\|y\| \le 1} \|Ay\|$$

for matrices in $C_{n\times n}$. Consider the product

$$R^{(M+\ell)} = \prod_{i=1}^{M+\ell} R_i$$

for $\ell \ge 1$. Then

$$\begin{aligned}
\|R^{(M+\ell)}\| &= \prod_{i=1}^{M+\ell} \|R_i\| = \prod_{i=1}^{M} \|R_i\| \quad \left\| \prod_{i=M+1}^{M+\ell} {}_i\tilde{R} \right\| \\
&\le n \prod_{i=1}^{M} \|R_i\| \max_{1\le j,k\le n} \left\{ \left| \left[\prod_{i=1}^{M+\ell} {}_i\tilde{R} \right]_{jk} \right| \right\} \\
&\le n \prod_{i=1}^{M} \|R_i\| \max_{1\le j,k\le n} \left\{ \left[\prod_{i=M+1}^{M+\ell} \left| {}_i\tilde{R} \right| \right]_{jk} \right\}
\end{aligned}$$

where, for any matrix B, $|B|$ is defined by

$$[\,|B|\,]_{jk} = |B_{jk}|.$$

However, for $i > M$, we have

$$|{}_i\tilde{R}| \le A_\xi = \begin{bmatrix} \xi & 1 & & & \\ & \xi & 1 & & \mathbf{0} \\ & & \xi & 1 & \\ & & & \ddots & \\ & \mathbf{0} & & \xi & 1 \\ & & & & \xi \end{bmatrix}$$

so

$$\left[\prod_{i=M+1}^{M+\ell} |{}_i\tilde{R}| \right]_{jk} \le \left[A_\xi^\ell \right]_{jk}$$

$$= \begin{cases} \binom{\ell}{k-j} \xi^{\ell-k+j}; & j \le k \le min\{n, l+j\}, \\ 0; & \text{otherwise,} \end{cases}$$

$$= \begin{cases} \xi^{\ell'-k+j} \binom{\ell'}{k-j} \prod_{r=\ell'}^{\ell} \frac{r\xi}{[r-(k-j)]}; & j \le k \le \min\{n, \ell+j\}, \\ 0; & \text{otherwise} \end{cases},$$

where $\ell' \le \ell$. Let $\xi < \eta < 1$ and choose ℓ' and ℓ large enough such that $\ell' \ge k-j$, ℓ' even and

$$\frac{r\xi}{[r-(k-j)]} < \eta$$

for all $r \ge \ell'$ and for all $j, k \in \{1, \ldots, n\}$. Then we have

$$\|R^{M+\ell'+r}\| \le n \left(\prod_{i=1}^{M} \|R_i\| \right) \binom{\ell'}{\ell'/2} \eta^r$$

so that

$$\|S\| \le \sum_{j=1}^{M+\ell'} \prod_{i=1}^{j} \|R_i\| + \frac{n}{1-\eta} \binom{\ell'}{\ell'/2} \prod_{i=1}^{M} \|R_i\|.$$

(ii) divergence

If $sp(R_i) \ge \xi > 1$ for $i \ge M$, then by an argument similar to that in (i) above, we can show that there exists an integer M' such that

$$\left\| \prod_{i=1}^{M'+\ell} R_i \right\| \ge \xi^{\ell},$$

for $\ell = 0, 1, 2, \ldots$. Thus the series diverges. □

CHAPTER THREE

RETRIAL QUEUES WITH FINITE BUFFERS

3.1 INTRODUCTION

So far, models of retrial queues with buffers have always assumed exponential service times and most have assumed Poisson arrivals as well. With the exception of Jonin and Sedol (1973) and Eldin (1967) described below, this is also true for multiserver retrial queues . Since multiserver queues have a structure similar to state dependent queues with buffers, we consider both cases together here. The M/M/s retrial queue was first considered by Wilkinson (1956). The first model of a retrial queue with buffer was studied by Kornishov (1974) and an extensive investigation was made by Rideout (1984). Yang (1990) investigated the GI/M/s/m retrial queue and Yang, Posner and Templeton (1992) developed numerical methods for the special case of Coxian interarrival times.

A multiserver model with primary service was considered by Jonin and Sedol (1973). In this model, a customer receives an exponentially distributed preliminary service (corresponding to connect time). If the connection is realized (with probability p) then the customer begins an exponentially distributed main service. Otherwise (with probability $1-p$) the customer joins a source of repeated calls. A variation was also considered by Eldin (1967) in which the future of a call is determined at the epoch of arrival and the call either proceeds through a short service (connect time only) and then joins the source of repeated calls or proceeds through a long service (connect time plus main service) and then leaves the system. Both of these models are similar to a multiserver retrial queue with two dimensional phase type service time

distributions (generalized Erlang for the first model and hyperexponential for the second model) except that the customer enters the orbit if he passes through phase one only, whereas in the M/PH_2/s retrial queue customers leave the system if they exit service from either phase. These models are fundamentally different from models without the preliminary service however because they allow customers to enter the orbit or source of repeated calls at times when there are free servers or waiting positions.

The fundamental property of retrial queues with buffers is the following: customers only enter the orbit when the buffer is full. This fact leads to substantial simplification and savings in computaional effort for Markov chain models of such systems. We model retrial queues with buffers as MAP(j)/SM(j)/1/b retrial queues where the index j represents the number of customers in the buffer. We model the service process as a state dependent MAP-like semi-Markov process. This allows us to include multiserver retrial queues since we can consider a MAP/PH/s queue as a MAP/SM(j)/1 queue with state dependent MAP-like service process. The dependence on the sub-level j also allows us to include models with overload control. We will not consider all of these models in detail but begin with the most general model to illustrate the fundamental property common to all of them. We make the following assumptions about the queue under consideration:

(a). The state of the queue at any time can be completely specified by the triplet (i, j, k) where i represents the number of customers in the orbit, $j \leq b$ represents the number of customers in service or waiting in a buffer position and k is a supplementary phase variable. The sojourn time in each state is exponentially distributed.

(b). Customers arriving to the queue enter service immediately if any servers are free or enter the FIFO buffer if all servers are busy but a

buffer position is free.

(c). A customer enters the orbit only if the buffer is full at the epoch of her arrival. This arrival may affect the phase k of the system but the buffer will remain full after she joins the orbit.

(d). Customers in the orbit retry for service periodically and enter the buffer if a position is free at the epoch of their retrial. Inter-retrial times for each customer are exponentially distributed with rate parameter θ. Entrance into the buffer may affect the phase k of the queue but i and j must decrease and increase by one respectively.

(e). Only one customer may arrive to the queue, complete service or enter the buffer from orbit at a time: i.e. no bulk service, arrivals or retrials.

(f). The arrival and service processes are independent of the number of customers in the orbit, although they may depend in an arbitrary way on the number of customers in the buffer.

Assumption (a) implies that the queue can be modelled as a continuous time Markov chain with state space $\sigma = \{(i,j,k)|i = 1,2,\ldots; j = 0,1,\ldots,b; k = 1,2,\ldots,m_j\}$. Assumption (e) guarantees that the generator of the Markov chain is block tridiagonal and that the diagonal blocks are themselves block tridiagonal. Assumption (c) implies that the superdiagonal blocks have only one nonzero block in the diagonal position corresponding to a full buffer. Assumption (d) implies that the subdiagonal blocks are block superdiagonal and assumption (f) gives the level dependence of the blocks a particularly simple form. The generator

has the form:

$$Q = \begin{bmatrix} B_0 & A_0 & & & \\ C_1 & B_1 & A_1 & & \\ & C_2 & B_2 & A_2 & \\ & & \ddots & \ddots & \ddots \end{bmatrix} \tag{3.1}$$

where the blocks A_i, B_i, and C_i have the general forms

$$A_i = \begin{bmatrix} 0 & & & \\ & \ddots & & \\ & & 0 & \\ & & & B_{b0} \end{bmatrix}$$

$$B_i = \begin{bmatrix} B_{01} - i\theta I & B_{00} & & & \\ B_{12} & B_{11} - i\theta I & B_{10} & & \\ \ddots & \ddots & \ddots & & \\ & B_{b-1,2} & B_{b-1,1} - i\theta I & B_{b-1,0} & \\ & & B_{b2} & B_{b1} & \end{bmatrix}$$

$$C_i = i\theta \begin{bmatrix} 0 & K_0 & & & \\ & \ddots & K_1 & & \\ & & \ddots & \ddots & \\ & & & 0 & K_{b-1} \\ & & & & 0 \end{bmatrix} \tag{3.2}$$

Assumption (d) implies that the matrices K_j are all stochastic.

The MAP/PH/1/b retrial queue is perhaps the most straightforward example of a queue which falls into the category described above. Consider a retrial queue with MAP arrival process represented by (D_0, D_1) and phase type service time distribution represented by (β^T, S). There is a single server, b buffer positions and retrial rate θ per customer. Then we have

$$B_{01} = D_0 \qquad B_{00} = D_1$$

$$B_{j1} = I \otimes S + D_0 \otimes I \qquad j = 1, \ldots, b$$

$$B_{j2} = S^0\beta^T \qquad K_j = I \qquad B_{j0} = B_0 = D_1 \otimes I.$$

Depending on the size of the buffer and the representations for the arrival and service processes, this model may prove too large to analyze in a reasonable time. If the dimensions of the arrival and service process representations are m and n respectively, then there are $m(1 + nb)$ states for each level (number of customers in the orbit) and the time complexity of the algorithm which calculates the stationary distribution is $O(Nbm^3n^3)$ where N is the level at which the system is truncated. The most obvious way to keep the complexity down is to keep m and n relatively small. For example, two dimensional representations are often sufficient if we only want to specify the first two or three moments of interarrival or service times (We can match the first two moments if the coefficient of variation is between .5 and 1 or the first three moments if the coefficient of variation is greater than 1). Models with exponential interarrival or service times (eg. M/PH/1, PH/M/1 retrial queues) where $m = 1$ or $n = 1$ may be sufficient for some applications.

The MAP/PH_2/s/b retrial also belongs to the class of queues considered here. Consider a retrial queue with s servers and $b - s$ buffer positions. The arrival process is a MAP with representation (T_0, T_1) and the service time for each server has a phase type distribution with two dimensional representation (β^T, S) where $\beta^T = (\beta_1, \beta_2)$ and

$$S = \begin{bmatrix} -(a_1 + g_1) & a_1 \\ a_2 & -(a_2 + g_2) \end{bmatrix}.$$

Interretrial times are exponentially distributed with rate θ per customer. The state space for the associated Markov chain is given by

$$\{(i, j, k, \ell) | i = 0, 1, \ldots; j = 0, \ldots, b; k = 1, \ldots, m; \ell = 0, \ldots, min\{j, s\}\}$$

. i represents the number in orbit, j the number in service or waiting in the buffer, k the arrival phase and ℓ the number of customers in phase 2

of thier service. If we order the state space with the labels (i, j, k, ℓ) in lexicographic order, the generator has the form in (3.1) and (3.2) with :

$$B_{j0} = \begin{cases} T_1 \otimes U_j & j = 0, \ldots, s-1 \\ T_1 \otimes I_{s+1} & j = s, \ldots, b \end{cases}$$

$$B_{j1} = \begin{cases} T_0 & j = 0 \\ I \otimes S_j + T_0 \otimes I_{j+1} & j = 1, \ldots, s \\ I \otimes S_s + T_0 \otimes I_{s+1} & j = s, \ldots, b \end{cases}$$

$$B_{j2} = \begin{cases} I \otimes D_j & j = 1, \ldots, s \\ I \otimes I_{s+1} & j = s+1, \ldots, b \end{cases}$$

$$K_j = I \otimes U_j$$

where D_j, S_j and U_j are given by

$$S_j = \begin{bmatrix} s_{j01} & s_{j00} & & \\ s_{j12} & \ddots & \ddots & \\ & \ddots & \ddots & s_{j,j-1,0} \\ & & s_{jj2} & s_{jj1} \end{bmatrix} \in \Re_{(j+1)\times(j+1)}$$

$$D_j = \begin{bmatrix} d_{j01} & & \\ d_{j02} & \ddots & \\ & \ddots & d_{j,j-1,1} \\ & & d_{jj2} \end{bmatrix} \in \Re_{(j+1)\times j}$$

$$U_i == \begin{bmatrix} \beta_1 & \beta_2 & & \\ & \ddots & \ddots & \\ & & \beta_1 & \beta_2 \end{bmatrix} \in \Re_{(j+1)\times(j+2)}.$$

I_j is the j-dimensional identity matrix and the scalars $s_{j\ell r}$ and $d_{j\ell r}$ are defined by

$$s_{j\ell 0} = (j-\ell)a_1 \qquad s_{j\ell 2} = \ell a_2 \qquad s_{j\ell 1} = -(\ell(a_2+g_2)+(j-\ell)(a_1+g_1))$$

$$d_{j\ell 2} = \ell g_2 \qquad d_{j\ell 1} = (j-\ell)g_1.$$

If ℓ customers are in phase 2 of their service, then $j-\ell$ are in phase 1. $s_{j\ell 0}$ is the rate of phase 1 customers switching to phase 2 and $s_{j\ell 2}$ is the rate of phase 2 customers switching to phase 1. $d_{j\ell 1}$ and $d_{j\ell 2}$ are the service completion rates of customers in phase 1 and phase 2 respectively. D_i has no superdiagonal elements because the number of customers in phase 2 cannot increase due to a service completion. This number will remain the same ($d_{j\ell 1}$) if a phase 1 customer completes service and will decrease by one ($d_{j\ell 2}$) if a phase 2 customer completes service. The form of U_i arises from the fact that the number of phase 2 customers will remain the same if a customer begins a service in phase 1 (with probability β_1) and will increase by one if a customer begins service in phase 2.

3.2 STABILITY CONDITION

We obtain a sufficient condition for ergodicity by applying Mustafa's criterion:

Proposition 1: *If the homogeneous quasi birth death process with B_{b0}, B_{b1} and $B_{b2}K_{b-1}$ for superdiagonal, diagonal and subdiagonal blocks respectively is irreducible and ergodic then so is the retrial queue with generator Q.*

Proof: Let P be the transition probability matrix for the jump chain of the retrial queue imbedded at each event and let P_b be the corresponding matrix for the homogeneous QBD process with blocks $A'_b = \Delta_b^{-1}B_{b0}$, $B'_b = I + \Delta_b^{-1}B_{b1}$ and $C'_b = \Delta_b^{-1}B_{b2}K_{b-1}$ where $\Delta_b = -diag(B_{b1})$. Let R be the rate matrix associated with this chain and let $\eta < 1$ be its spectral radius. From the discussion in the proof of Lemma 1.3.4 in Neuts (1981), we know that for all $\epsilon \in (\eta, 1)$, $sp(A'_b +$

$\epsilon B'_b + \epsilon^2 C'_b) = \epsilon' < \epsilon$. Let $x_b > 0$ be the right eigenvector associated with ϵ' so that $(A'_b + \epsilon B'_b + \epsilon^2 C'_b)x_b = \epsilon' x_b < \epsilon x_b$. We now define a vector $f = [f_0, f_1, \ldots]$ with $f_i = [f_{i0}, f_{i1}, \ldots, f_{ib}]$ such that, for $i > 1$,

$$f_{ij} = \epsilon^{-i}(ae + \epsilon^{b-j} K_j K_{j+1} \ldots K_{b-1} x_b)$$

where $a \in (0,1)$. To prove ergodicity, we need to show that $(P - I)f < -\epsilon'' e$ (except for a finite number of components) for some $\epsilon'' > 0$. Now for $j = 0, \ldots, b-1$ we have

$$[(P - i)f]_{ij} = \epsilon^{-i}(i\theta I + \Delta_j)^{-1}[(\epsilon^2 B_{j2} K_{j-1} K_j + \epsilon B_{j1} K_j + B_{j0}))K_{j+1} \ldots K_{b-1} x_b - i\theta a(1-\epsilon)]$$

where $\Delta_j = -diag(B_{j1})$. For $j = b$, we have

$$[(P - I)f]_{ib} = \epsilon^{-(i+1)}[a(1-\epsilon)A'_b e + (A'_b + \epsilon B'_b + \epsilon^2 C'_b - \epsilon I)x_b]$$

. Since $\beta_1 = -(A'_b + \epsilon B'_b + \epsilon^2 C'_{ib} - \epsilon I)x_b > 0$, we can choose $a \in (0,1)$ so that $[(P - I)f]_{ib} < -\epsilon'' e = -min\{\beta_1\}/2$. Since $[(P - I)f]_{ij} \to -\infty$ as $i \to \infty$ for $j = 0, \ldots, b-1$, there exists an integer i_1 such that $[(P - I)f]_{ij} < -\epsilon'' e$ for all $i > i_1$. □

A necessary and sufficient condition for the ergodicity of the homogeneous chain associated with P_b can be obtained by applying Theorem 1.7.1 in Neuts (1981). The condition is given by $\pi(B_{b0} - B_{b2}K_{b-1})e < 0$ where π is the solution to $\pi(B_{b0} + B_{b1} + B_{b2}K_{b-1}) = 0$ and $\pi e = 1$. It is interesting to note that the sufficient condition in Proposition 1 depends only on the behavior of the queue when the buffer is full. This is similar to the well known condition $\lambda/s\mu < 1$ for ergodicity of the $M/M/s$ queue since $s\mu$ is the service rate when all service positions are occupied. In both cases, the ergodicity condition takes this form not because the full buffer state corresponds to the fastest service but

because the queue must pass through this state in order to add any customers to the queue (or orbit).

3.3 SOLUTION METHOD

The sparsity of the A_i together with the block tridiagonal form of the B_i allow a substantial saving in the computational effort required to calculate the stationary distribution of states and the distribution of waiting time for the queue. As in the previous chapters, this requires the solution of systems of the form $yQ = \alpha$. For example, the stationary distribution x is the solution of the system $xQ = 0$. In all other cases, α will be non-zero and Q will be a subgenerator with a form similar to that in (3.1)-(3.2). In that case, for efficiency of notation, we will still denote the subgenerator by Q and maintain the same labels for its various components (A_i, B_{j1} etc.). We proceed, as in Chapter One , by eliminating levels one at a time via block Gaussian elimination, starting from the lowest level. The basic algorithm used is the same as Algorithm 1.1, but we repeat the description of this algorithm below, with the notation used in this chapter.

Algorithm 3.1

Reduction Phase

$i \leftarrow 0$

$B'_0 \leftarrow B_0$

$\alpha'_0 \leftarrow \alpha_0$

Until $i = N$, do

$B'_i \leftarrow B_{i1} - C_i(B'_{i-1})^{-1}A_{i-1}$

$\alpha'_i \leftarrow \alpha_i - \alpha'_{i-1}(B'_{i-1})^{-1}A_{i-1}$

$i \leftarrow i + 1$

End

Middle Phase

Solve $[y_N, y_{N+1}, \ldots]Q_N = [\alpha'_N, \alpha_{N+1}, \ldots]$ where Q_N is the generator obtained from Q by eliminating all levels below N and replacing B_N with B'_N.

Expansion Phase

$i \leftarrow N - 1$

Until $i = 0$, do

$y_i \leftarrow (\alpha_i - y_{i+1}C_{i+1})(B'_i)^{-1}$

$i \leftarrow i - 1$

End

In order to accomplish the middle phase of this algorithm, it is necessary to truncate or approximate the generator in some way so that we can eliminate the level dependence above some level N. We do so by making the approximating assumption that when there are more than N customers in the orbit, one customer from the orbit joins the buffer immediately after each service completion. Then the buffer is always full when there are more than N customers in the orbit and the approximate generator has the form:

$$Q = \begin{bmatrix} B_0 & A_0 & & & & & & \\ C_1 & B_1 & A_1 & & & & & \\ & \ddots & \ddots & \ddots & & & & \\ & & C_N & B_N & A_N & & & \\ & & & C_{N+1} & B & A & & \\ & & & & C & B & A & \\ & & & & & \ddots & \ddots & \ddots \end{bmatrix}$$

where $A = B_{b0}$, $B = B_{b1}$, $C = B_{b2}K_{b-1}$, $C_{N+1} = [0, \ldots, 0, C]$ and

$$A_N = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ A \end{bmatrix}.$$

If we are calculating the stationary distribution x, we can accomplish the middle phase of algorithm 3.1 by solving the system

$$x_{N+1}[B + RC - C_{N+1}(B'_N)^{-1}A_N] = 0$$

where R is the minimal nonnegative solution to the matrix quadratic equation $A + RB + R^2C = 0$ which can be obtained by applying the methods in Neuts (1981). We can then obtain $x_N, x_{N-1}, \ldots, x_0$ by applying the expansion phase of Algorithm 3.1 (with $y \leftarrow x$ and $\alpha = 0$) and the components of x above level N are given by $x_{N+1+j} = x_{N+1}R^j$. We must normalize x so that $xe = x_{N+1}(I - R)^{-1}e + \sum_{i=0}^{N} x_i e = 1$. In practice, we gradually increase the value of N until the probability $x_{N+1}(I - R)^{-1}e$ of being above level N is sufficiently small (perhaps 10^{-5} or so). We then truncate the system by replacing B with $B + RC$ and deleting all levels above $N + 1$. In all further calculations, when we solve systems of the form $yQ = \alpha$, Q will be a subgenerator obtained by altering this finite generator in some way and the system $yQ = \alpha$ will be finite. In that case, the middle phase corresponds to solving $y_{N+1}[B + RC - C_{N+1}(B'_N)^{-1}A_N] = \alpha'_{N+1}$.

3.4 IMPLEMENTING THE ALGORITHM

Consider the matrices $\hat{G}_i = -(B'_i)^{-1}A_i$ for $i = 0, 1, \ldots$ where the B'_i are as defined in Algorithm 3.1 above. When Q is a generator, the (s, t)th element of $\hat{G}_i$ represents the probability that, starting from state

(i, s), the process eventually reaches level $i+1$ and enters it at phase t (Here "phase "refers to number j in buffer as well as phase k). $\hat{G}_i$ is the dual of the matrix G_i defined in Pearce (1994) in the same sense that $\hat{R}_i$ defined in Chapter 2 is the dual of R_i defined in Pearce (1994) and Bright and Taylor (1995). We get one from the other by reversing the order of levels and a proof for the interpretation of and $\hat{G}_i$ would follow exactly the corresponding proof for G_i which appears in Pearce (1994).

The reduction phase of Algorithm 3.1 can be expressed as the recursions

$$\hat{G}_i = -(B_i + C_i \hat{G}_{i-1})^{-1} A_i$$

or equivalently

$$A_i + B_i \hat{G}_i + C_i \hat{G}_{i-1} \hat{G}_i = 0. \tag{3.3}$$

and

$$\alpha'_i = \alpha_i + \alpha'_{i-1} \hat{G}_{i-1} \tag{3.4}$$

From the form of A_i and from the interpretation of $\hat{G}_i$ it is clear that $\hat{G}_i$ must have the form

$$\hat{G}_i = \begin{bmatrix} 0 & \dots & 0 & \hat{G}_{i0} \\ 0 & \dots & 0 & \hat{G}_{i1} \\ \vdots & & \vdots & \vdots \\ 0 & \dots & 0 & \hat{G}_{ib} \end{bmatrix}.$$

. If we substitute this form and the forms of the blocks of the generators

into (3) we obtain a system of the form

$$\begin{bmatrix} B_{01}-i\theta I & B_{00} & & & i\theta K_0\hat{G}_{i-1,1} \\ B_{12} & B_{11}-i\theta I & B_{10} & & \vdots \\ & \ddots & \ddots & \ddots & i\theta K_{b-2}\hat{G}_{i-1,b-1} \\ & & & & B_{b-1,0}+i\theta K_{b-1}\hat{G}_{i-1,b} \\ & & & B_{b2} & B_{ib1} \end{bmatrix} \times \begin{bmatrix} \hat{G}_{i0} \\ \vdots \\ \vdots \\ \vdots \\ \hat{G}_{ib} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ -B_{b0} \end{bmatrix} \tag{3.5}$$

. The system in (3.5) is tridiagonal except for the last column. We can thus solve the system efficiently (in O(b) time) by using block Gaussian elimination, eliminating one sublevel at a time, starting from the lowest sublevel. Each step is obtained according to the general model implied by (1.3) and (1.4) in Chapter One except that we have a system with column vectors instead of row vectors. The analogs of these equations are

$$Q_t^* y_t = \alpha_t^* \qquad \text{and} \qquad y_s = Q_s^{-1}(\alpha_s' - Q_{st}y_t) \tag{3.5}$$

where

$$Q_t^* = Q_t - Q_{ts}Q_s^{-1}Q_{st} \qquad \text{and} \qquad \alpha_t^* = \alpha_t - Q_{ts}Q_s^{-1}\alpha_s. \tag{3.6).}$$

If we eliminate the lower sublevels first, after the jth stage we obtain a

system of the form

$$\begin{bmatrix} B'_{ij1} & B_{j0} & & & H_{ij} \\ B_{j+1,2} & \ddots & \ddots & & \vdots \\ & \ddots & \ddots & \ddots & i\theta K_{b-2}\hat{G}_{i-1,b-1} \\ & & \ddots & & B_{b-1,0}+i\theta K_{b-1}\hat{G}_{i-1,b} \\ & & & B_{b2} & B_{b1} \end{bmatrix} \times \begin{bmatrix} \hat{G}_{ij} \\ \vdots \\ \vdots \\ \vdots \\ \hat{G}_{ib} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ -B_{b0} \end{bmatrix} \tag{3.6}$$

where the blocks B'_{ij1} and H_{ij} are defined by the following algorithm which solves the system of equations:

Algorithm 3.2

Reduction Phase

$B'_{i01} \leftarrow B_{01} - i\theta I$
$H_{i0} \leftarrow i\theta K_0 \hat{G}_{i-1,1}$
$j \leftarrow 1$
Until $j = b-1$, do
$B'_{ij1} \leftarrow B_{j1} - i\theta I - B_{j2}(B'_{i,j-1,1})^{-1}B_{j-1,0}$
$H_{ij} \leftarrow i\theta K_j \hat{G}_{i-1,j+1} - B_{j2}(B'_{i,j-1,1})^{-1}H_{i,j-1}$
$j \leftarrow j+1$
End
$H_{i,b-1} \leftarrow B_{b-1,0} + i\theta K_{b-1}\hat{G}_{i-1,b} - B_{b-1,2}(B'_{i,b-2,1})^{-1}H_{i,b-2}$
$B'_{ib1} \leftarrow B_{b1} - B_{b2}(B_{i,b-1,1})^{-1}H_{i,b-1}$

Middle Phase

$$\hat{G}_{ib} \leftarrow -(B'_{ib1})^{-1} B_{b0}.$$

Expansion Phase

$\hat{G}_{i,b-1} \leftarrow -(B'_{i,b-1,1})^{-1} H_{i,b-1} \hat{G}_{ib}$
$j \leftarrow b-2$
Until $j = 0$, do
$\quad \hat{G}_{ij} \leftarrow -(B'_{ij1})^{-1}(B_{j0}\hat{G}_{i,j+1} + H_{ij}\hat{G}_{ib})$
End

Algorithm 3.2 provides the components of $\hat{G}_i$ for $i = 0, 1, \ldots, N$ and (3.4) provides α'_{N+1}. With the approximate generator we have

$$\hat{G}_N = \begin{bmatrix} \hat{G}_{N0} \\ \vdots \\ \hat{G}_{Nb} \end{bmatrix}$$

i.e. the large blocks of zeros are missing because level N has only one sublevel. The middle phase of algorithm 3.1 provides y_{N+1} according to

$$y_{N+1}[B + RC + C\hat{G}_{Nb}] = \alpha'_{N+1}.$$

The expansion phase of Algorithm 3.1 provides $y_N, \ldots, y_0$ according to

$$y_i[B_i + C_i\hat{G}_{i-1}] = \alpha'_i - y_{i+1}C_{i+1} \tag{3.7}$$

. Fortunately the matrix $B_i + C_i\hat{G}_i$ is the same as the square matrix in (3.5) so we can use use the same algorithm to solve this system without repeating the steps necessary to reduce the matrix. The following algorithm summarizes this method:

Algorithm 3.3

(To solve: $y_i[B_i + C_i\hat{G}_{i-1}] = \gamma_i$ for y_i)

Reduction Phase

$j \leftarrow 0$

$\gamma'_{i0} \leftarrow \gamma_{i0}$

$\gamma'_{ib} \leftarrow \gamma_{ib}$

Until $j = b-1$, do

$\gamma'_{ij} \leftarrow \gamma_{ij} - \gamma'_{i,j-1}(B'_{i,j-1,1})^{-1}B_{j-1,0}$

$\gamma'_{ib} \leftarrow \gamma_{ib} - \gamma'_{i,j-1}(B'_{i,j-1,1})^{-1}H_{i,j-1}$

$j \leftarrow j+1$

End

$\gamma'_{ib} \leftarrow \gamma_{ib} - \gamma'_{i,b-1}(B'_{i,b-1,1})^{-1}H_{i,b-1}$

Middle Phase

$y_{ib} \leftarrow \gamma'_{ib}(B'_{ib})^{-1}$

Expansion Phase

$j \leftarrow b-1$

Until $j = 0$, do

$y_{ij} \leftarrow (\gamma_{ij} - y_{i,j+1}B_{j+1,2})(B'_{ij1})^{-1}$

$j \leftarrow j-1$

End

where $\gamma_i = [\gamma_{i0}, \ldots, \gamma_{ib}]$. We can use Algorithm 3.3 to solve (3.7) if we set $\gamma_{i0} = \alpha_{i0}$ and $\gamma_{ij} = \alpha_{ij} - i\theta y_{i+1,j-1}K_{j-1}$ for $j = 1, \ldots, b$.

If the service time for a MAP/PH/1/b retrial queue follows an

Erlang or generalized Erlang distribution, the reduction phase of Algorithm 3.2 can be performed in $O(bm^3n)$ time. This is because the matrices B_{j2} are non-zero only in the last row of blocks corresponding to service completion from the last stage of the generalized Erlang. Since the matrices B_{j1} are also block tridiagonal, the form of the B'_{ij1} are block tridiagonal except for the last row of blocks. The system of equations associated with evaluating $B_{j2}(B'_{i,j-1,1})^{-1}$ then has a form similar to that in (3.4) and Algorithm 3.3 can be adapted to solve the system. Similarly, if the arrival process is Erlang or generalized Erlang renewal, then only the first column of blocks of $\hat{G}_{ij}$ are non-zero. This is because the chain always enters the next highest level via the first arrival phase (recall the interpretation of $\hat{G}_i$). Because of this, both the reduction phase and the expansion phase of Algorithm 3.2 can be executed in $O(bmn^3)$ time.

The number of states for the MAP/PH$_2$/s/b retrial model is $Nm(s+1)(b-s/2)$ and the time complexity of finding the stationary distribution is $O(Nbm^3s^3)$. As in the previous example, this can be reduced to $O(Nbms^3)$ if the arrival process is Erlang or generalized Erlang renewal.

3.5 WAITING TIME FOR MAP/PH/1/b RETRIAL QUEUE

In a retrial queue with buffer, the waiting time consists of the time in orbit as well as the time in the buffer. The distribution and moments of the waiting time in orbit can be obtained using the methods of Sections 1.8 and 1.9. Perhaps of more interest is the total waiting time: i.e. the sum of the waiting times in the orbit and in the buffer. Intuitively, we expect the addition of a buffer to have a small effect on the mean waiting time due to the elimination of server idle time

when there are customers in the orbit. We expect the variability of the total waiting time to decrease with the addition of a buffer since the essentially random service order of the orbit becomes less important. This reduction in waiting time variability may be the motivation for introducing a buffer into a retrial queue.

For simplicity, we consider the total waiting time of a MAP/PH/1/b retrial queue, although the MAP/PH_2/s/b can also be examined using the same method. We choose an arbitrary arrival of a customer, tag that customer, and measure his total waiting time. When the tagged customer is in orbit, we must keep track of the number of customers in orbit, the number of customers in the buffer, and the phase of the arrival and service processes. Once the tagged customer enters the buffer, we only need to keep track of his position in the buffer and the phase of the current service. This is because the waiting time in buffer of a customer is not affected by the presence of customers in the orbit or by the presence in the buffer of customers which arrive after him. This is true for the MAP/PH/1/b and MAP/PH_2/s/b retrial queues although it is not true for retrial queues with overload control where the service times depend on the number of customers in the buffer. In that case, we would need to keep track of the arrival phase and the number of customers in the orbit and buffer as well. We thus construct an absorbing Markov chain on the state space $\Sigma = \sigma \cup \sigma_{buffer} \cup \{s\}$ where σ is the state space defined in Section 3.1, $\{s\}$ is the absorbing state corresponding to the tagged customer having entered service and $\sigma_{buffer} = \{(j,k)|j = 1,\ldots,b-1; k = 1,\ldots,n\}$ contains the states with the tagged customer in the buffer. The parameter j in σ_{buffer} reresents the number of customers in the buffer; not the number of customers in service and in the buffer. The generator for this absorbing Markov

chain has the form

$$Q' = \begin{bmatrix} Q_o & Q_{ob} & Q_{os} \\ 0 & Q_b & Q_{bs} \\ 0 & 0 & 0 \end{bmatrix}.$$

Define P_f to be the $b+1$ dimensional diagonal matrix with a one in the last diagonal position and and zeros everywhere else: The one corresponds to a full buffer. Then $Q_o = Q - \theta I \otimes (I - P_f) \otimes I \otimes I$ (where Q is the generator in (3.1)), corresponds to the tagged customer in orbit. $Q_{ob} = \theta e \otimes J_{ob} \otimes e \otimes I$ corresponds to the tagged customer entering the buffer where

$$J_{ob} = \begin{bmatrix} 0 \\ I_{b-1} \\ 0 \end{bmatrix}$$

and I_{b-1} is the $b-1$ dimensional identity matrix. The portion of the generator corresponding to the tagged customer already in the buffer is

$$Q_b = \begin{bmatrix} S & & & \\ S_0\beta^T & S & & \\ & \ddots & \ddots & \\ & & S^0\beta^T & S \end{bmatrix} \in \Re_{n(b-1)\otimes n(b-1)}.$$

$Q_{os} = -(Q_o + Q_{ob})e$ and $Q_{bs} = -Q_b e$ are column vectors corresponding to the tagged customer entering service (absorption) from orbit or from the buffer respectively.

The first step in obtaining the distribution or moments of the waiting time is to determine the initial vector associated with this absorbing chain: i.e. the stationary distribution of states after an arbitrary arrival. To do this, we must consider the absorbing chain with generator

$$Q'' = \begin{bmatrix} Q_{na} & Q_{orbit} & Q_{buffer} & Q_{service} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

where

$$Q_{orbit} = \begin{bmatrix} 0 & P_f \otimes D_1 \otimes I & & \\ & \ddots & P_f \otimes D_1 \otimes I & \\ & & \ddots & \ddots \end{bmatrix}$$

$$Q_{buffer} = e \otimes \begin{bmatrix} 0 \\ I_{b-1} \otimes D_1 \otimes I \\ 0 \end{bmatrix},$$

$$Q_{service} = e \otimes \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \otimes D_1 e$$

and Q_{na} is the subgenerator obtained from Q in (3.1) and (3.2) by making the substitutions $B_{j0} \leftarrow 0$, i.e. by turning off the arrivals. The stationary distribution of states at an arbitrary arrival is given by $x_a = xQ_{na}/xQ_{na}e$ where x is the stationary distribution ($xQ = 0$) of states at an arbitrary point in time. The stationary distribution just before an arbitrary arrival is $-x_a Q_{na}^{-1}$ and the distribution of states after the arrival of a tagged customer is

$$x_{tag} = \frac{-x \begin{bmatrix} Q_{orbit} & Q_{buffer} & Q_{service} \end{bmatrix}}{xQ_{na}e}.$$

The waiting time has a phase type distribution with representation (x_{tag}, Q). As in Section 1.8, we can use the randomization method to obtain the distribution of the waiting time according to $F(t) = 1 - x_{tag} exp[Q'_s t]e$ where the subgenerator Q'_s is the upper left hand corner (delete absorbing state $\{s\}$) of Q':

$$Q'_s = \begin{bmatrix} Q_o & Q_{ob} \\ 0 & Q_b \end{bmatrix}.$$

Calculating the entire distribution in this manner will be quite time consuming although simple to implement. The jth moment of the waiting time is given by

$$M_j = (-1)^j j! x_{tag} (Q'_s)^{-j} e$$

and the inverse $(Q'_s)^{-1}$ is given by

$$(Q_s)^{-1} = \begin{bmatrix} Q_o^{-1} & -Q_o^{-1}Q_{ob}Q_b^{-1} \\ 0 & Q_b^{-1} \end{bmatrix}.$$

We can calculate the effect of post-multiplying a vector by Q_o^{-1} if we apply Algorithms 3.1,3.2 and 3.3 with $B_{j0} \leftarrow 0$. The effect of post-multiplying by Q_b^{-1} can be calculated easily using block Gaussian elimination and no reduction phase is required because the superdiagonal blocks are zero. In fact, because Q_b is homogeneous, this could be done quicker using Ye and Li's (1991) folding algorithm although the savings would be minimal compared to the effort required to postmultiply by Q_o^{-1}. We calculated the mean and the coefficient of variation of the total waiting time for the example retrial queue with various buffer sizes b. The results are illustrated in Figure 3.1 and Figure 3.2. Note the dramatic drop in the mean waiting time which results from adding the first buffer position ($b = 1 \rightarrow b = 2$). The plot of the coefficient of variation falls quite sharply as we go from $b = 2$ to $b = 1$. This is due to the dramatic change in the mean and not due to a radical change in the variance. In fact the curve for the variance appears monotonic. Clearly, for this example there is very little benefit to the mean in adding more than about 4 or 5 buffer positions, while the coefficient of variation continues to decrease until around $b = 9$ or $b = 10$,

Figure 3.1: Mean Waiting Time vs. Buffer Size

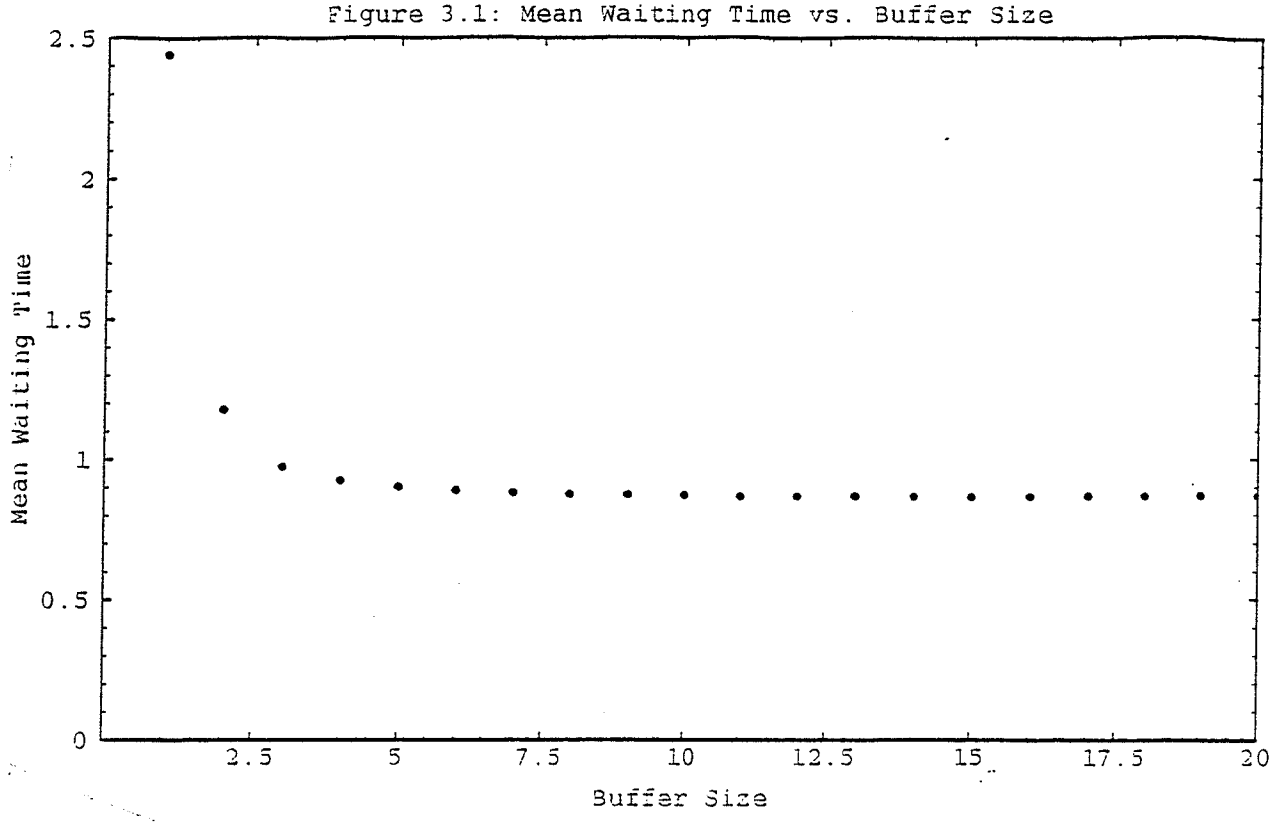
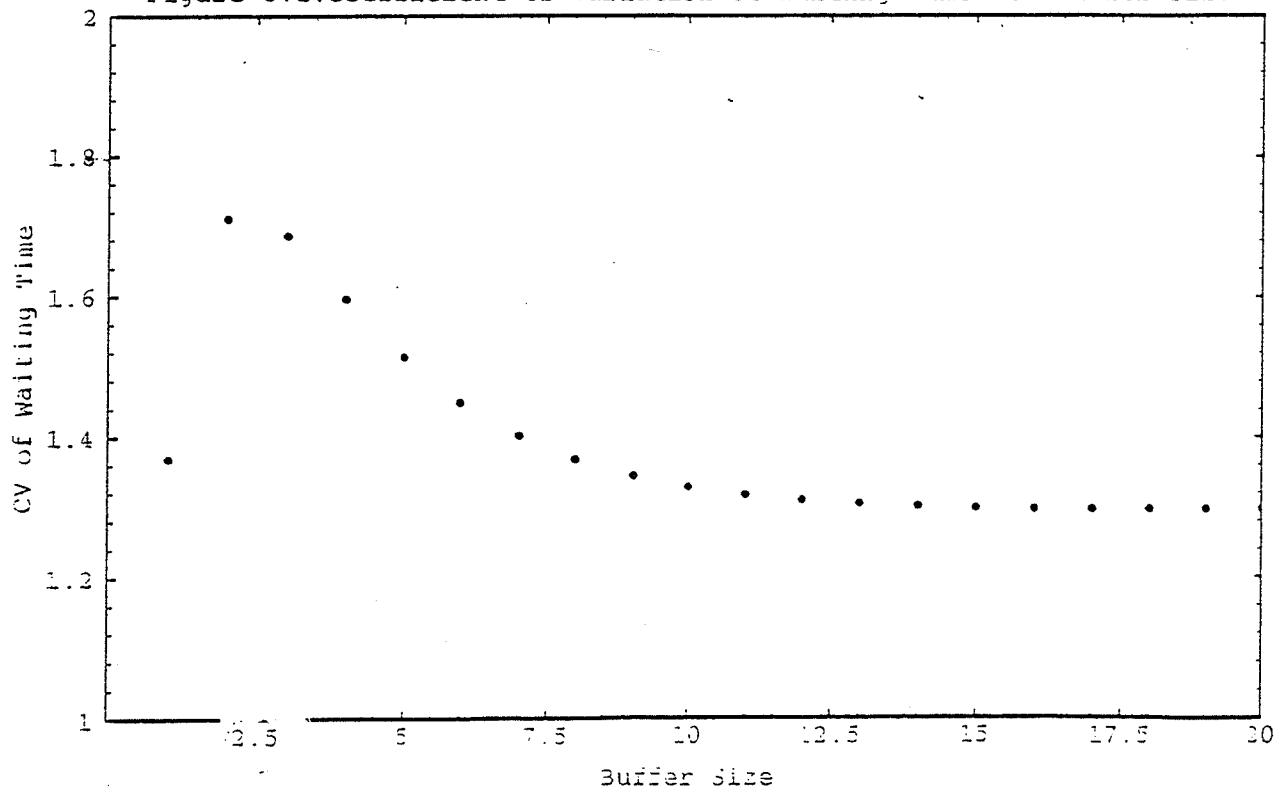


Figure 3.2:Coefficient of Variation of Waiting Time vs. Buffer Size



3.6 EXTENSION: FINITE NUMBER OF CUSTOMERS

The form of the generator given in (1) and (2) does not encompass models with a finite number of customers such as, for example, models of local area networks. In that case, the number of sublevels (b in the model above) varies with the number of customers in the orbit. If there are N customers in total, there can be at most $N-i$ customers in service or waiting in the buffer when there are i customers in the orbit. Also, the arrival process will depend on i the number of customers in the orbit as well as j the number of customers in service or waiting in the buffer since the number of active sources contributing to the input stream is $N-i-j$. For this model, the components B_{j0}, B_{j1} etc. of the generator become level dependent and so require an extra index (B_{ij0},B_{ij1} etc.). Algorithms 3.1, 3.2 and 3.3 however can be applied in the same way as above if we simply add the appropriate level indices to the components where required. The algorithms depend on the structure of the blocks of the generator and not the content of the components.

Consider an M/PH/1/b/N retrial queue with arrival rate λ for each of N customers, 1 server with service time distribution represented by $\beta^T, S)$, b buffer positions and exponential retrial times with rate θ per customer. The states are labeled by (i, j, k) respectively the number of customers in orbit, number of customers in service or waiting in the buffer and phase of service. Each level i has $b_i = min\{b, N-i\}$ sublevels and the components of the generator are given by:

$$B_{ij0} = \begin{cases} (N-i)\lambda\beta^T & j = 0 \\ (N-i-j)\lambda I & j = 1, \ldots, b_i \end{cases}$$

$$B_{ij1} = \begin{cases} -(N-i)\lambda & j = 0 \\ S - (N-i-j)I & j = 1, \ldots, b_i \end{cases}$$

$$B_{ij2} = \begin{cases} S^0 & j = 0 \\ S^0\beta^T & j = 1, \ldots, b_i \end{cases}$$

$$K_{ij} = I \qquad j = 0, \ldots, b_i.$$

We could also generalize this model by allowing the interarrival time for each customer to have a PH_2 type distribution (β^T, S) as defined in the previous section. This model without buffer is considered in detail in Chapter 4. The number of states for this model can be quite large so, in practice, the size of the service time representation will have to be kept small in order to avoid very lengthy computation times. The time required in the reduction phase can be reduced if the service time distribution is of Erlang type but the expansion phase is not accelerated. If (γ^T, T) denotes the service time representation, then the components of the generator are given by:

$$B_{ij0} = \begin{cases} D_{N-i} \otimes \gamma^T & j = 0 \\ D_{N-i-j} \otimes I & j = 1, \ldots, b_i \end{cases}$$

$$B_{ij1} = \begin{cases} S_{N-i} & j = 0 \\ I_{N-i-j+1} \otimes T + S_{N-i-j} \otimes I & j = 1, \ldots, b_i \end{cases}$$

$$B_{ij2} = \begin{cases} U_{N-j-1} \otimes T^0 & j = 0 \\ U_{N-j-1} \otimes T^0 \beta^T & j = 1, \ldots, b_i \end{cases}$$

$$K_{ij} = I_{N-i-j+1} \otimes I \quad j = 0, \ldots, b_i.$$

Other models with a finite number of customers include the M/PH_2/s/b/N and PH_2/PH_2/s/b/N retrial queues. We do not describe these in any detail since the previous examples are sufficient to demonstrate the construction of the generator for these and other similar models. We can also add overload control to any of these models by making the service time distribution depend on the number of customers in the buffer in some way. All of these models have the same fundamental structure so that the same or similar computational schemes can be applied. The only restriction is the computational time available which restricts the size of the models which can be utilized.

3.7 CONCLUSION

We have provided a computational scheme which can be used to determine the stationary distribution of states, the distributions of the waiting time and number of retrials and the moments of the waiting time for retrial queues with buffers. We have generalized significantly past models with exponential interarrival and service time distributions. The computational methods described can be applied to a wide range of models and computational complexity for large models is the limiting factor in the analysis of these models.

CHAPTER FOUR

APPROXIMATION METHOD FOR RETRIAL QUEUES WITH PHASE TYPE RETRIAL TIMES

4.1 INTRODUCTION

In almost all models of retrial queues, the time between retrials for any customer is assumed to be exponentially distributed. Kapyrin (1977) attempted to derive an analytic solution for the M/G/1 retrial queue with general retrial time distribution but the method and results were found to be incorrect (Falin, 1986). Pourbabai developed an iterative method for a G/M/K/0 queueing loss system with retrials in which the net retrial process from returning customers is approximated by matching parameters of the overflow process. Liang and Kulkarni (1993) introduced a relation, which they called K-dom inance on the class of phase type distributions and showed that, for single serve r retrial queues with general arrival processes and service time distributions an d phase type retrial times, longer retrial times with respect to this relation re sult in more congested systems in the stochastic sense.

The inherent difficulty with non-exponential retrial times stems from the fact that the model must, in some way, keep track of the elapsed retrial time for each of possibly a very large number of customers. The net retrial process from all customers is an extremely complex non-renewal process. Choi (1993) avoided this problem by considering an M/M/1 retrial queue with general retrial times where only one customer may attempt retrials from orbit. Liang (1991) developed an approximate method for obtaining the stationary distribution of queue size for M/G/1 retrial queues with retrial time distributions which are mixtures of Erlangs. Yang et al (1994), developed an effective approximation for

the M/G/1 retrial queue with general retrial times by noting that, for most applications, retrial times are significantly shorter than service times. In aa typical retrial queue, most customers in the orbit will make numerous requests during any given service interval. Thus, while the elapsed retrial times for different customers in orbit are dependent, the dependence will be very weak. The approximation assumes that the elapsed retrial time for any customer is a random variable independent of other customers elapsed retrial times. According to renewal theory, the distribution of elapsed retrial time observed at a point far from the time origin has the form $m(x) = \int_0^x \theta(1 - T(u))du$ (see Ross, 1983) where $T(u)$ is the cumulative distribution function, with mean $1/\theta$, of the retrial time. The approximation is used to derive the distribution of the number of customers in orbit and the mean waiting time and number of retrials per customer. The algorithm is shown to perform well in predicting the mean and variance of queue length by comparing to an exact solution for retrial queues with hyperexponential or Erlang distributions of order two for retrial and service times.

In this chapter we develop an approximation for retrial queues with phase type retrial time distributions. Our goal is not to improve upon the approximation used by Yang et al, but to extend the basic idea of the approximation so that we can use it to approximate higher moments or the entire distribution of the waiting time . Instead of using the approximation only in the calculation of the stationary distribution, as in Yang et al, we use it to approximate the generator of the queue itself. We can then use the numerical methods developed in previous chapters to obtain information about the waiting time from the approximate generator.

The use of numerical methods also makes it easier to extend the approximation to other models such as the M/M/s or MAP/Ph/1 retrial queue with phase type retrial times. We will only explore the M/PH/1

case in detail but the extensions to these other models are not difficult, except for possibly high dimensionality resulting in long computational times.

Although Yang et al. provide a necessary condition for stability, they do not show that the condition is also sufficient. We show that this is the case for phase type retrial times and extend the result to the multiserver case as well.

4.2 STABILITY CONDITION

The following proposition provides a necessary and sufficient condition for ergodicity of an M/G/1 retrial queue with phase type retrial times by considering the mean drift.

Proposition 4.1: *An M/G/1 retrial queue with phase type retrial times is ergodic if and only if* $\rho = \lambda/\mu < 1$ *where* λ *is the arrival rate to the queue and* μ^{-1} *is the mean service time.*

Proof: Assume that the retrial time distribution is of phase type with n dimensional representation (α^T, T) and consider the Markov chain embedded at the epochs of service completion. Let η_t denote the epoch of the tth service completion and let I_t denote the number of customers in orbit at time η_t^+. Then we have

$$I_{t+1} = I_t - B_t + Y_{t+1}$$

where $B_t \in \{0, 1\}$ is the number of orbiting customers who begin service in the interval (η_t, η_{t+1}) (i.e $B_t = 1$ if the $t+1$st customer served comes from the orbit and $B_t = 0$ if he comes from the primary arrival stream) and $Y_(t + 1)$ is the number of customers arriving during the $t + 1$st service. Clearly Y_t has mean ρ for all $t = 1, 2, \ldots$. Let the vector

$J_t = (j_1, \ldots, j_n)$ contain the numbers of customers in each of the n phases of retrial at time η_t^+ and let

$$p_k = e_k^T(\lambda I - T)^{-1}e,$$

where e_k is the indicator vector for phase k of the retrial time distribution, denote the probability that a customer in phase k makes no retrial before the next arrival of a primary customer. Then the mean drift of the embedded chain from state (i, J) is given by

$$\begin{aligned}
d_{(i,J)} &= E(I_{t+1} - I_t | I_t = i \quad and \quad J_t = (j_1, \ldots, j_n)) \\
&= E(Y_{t+1}) - E(B_t | I_t = i \quad and \quad J_t = J = (j_1, \ldots, j_n)) \\
&= \rho - 1 + P(B_t = 0 | I_t = i \quad and \quad J_t = (j_1, \ldots, j_n)) \\
&= \rho - 1 + \prod_{k=1}^{n} (p_k)^{j_k} \\
&\leq \rho - 1 + (\max_k \{p_k\})^i \to \rho - 1 \quad as \quad i \to \infty
\end{aligned}$$

Thus, if $\rho < 1$, the mean drift is negative for large enough i and the chain is ergodic. Yang et al (1994) have already shown that this is a necessary condition. □

The following generalizes Proposition 4.1to the case of multiserver queues although the conditions are more difficult to verify. Consider an M/PH/s retrial queue with phase type retrial times represented by (α^T, T) and service times by (β^T, S). Let $\sigma = \{1, \ldots, m\}$ denote the set of service phases and let $\Sigma = \sigma^s$ denote the state space of possible combinations of phases for all s servers. Let S_j be a random variable representing the service time for a service which starts in phase $j \in \sigma$. S_j has the cumulative distribution $F_j(t) = 1 - e_j^T exp[St]e$ where e_j is a column vector with a 1 in the jth position and 0 everywhere else. If $J = (j_1, j_2, \ldots, j_s) \in \Sigma$ is a vector recording the phases of all s

servers, then the cumulative distribution of the time τ_J until the next service completion is given by $G_J(t) = 1 - \prod_{k=1}^{s}(1 - F_{j_k}(t))$. Let $\mu = \max_{J\in\Sigma} E(\tau_J)$ denote the maximum expected time (over all possible server configrations) until the next service completion. Then we have the following.

Proposition 4.2: *An M/PH/s retrial queue with phase type retrial times is ergodic if $\rho = \lambda/\mu < 1$ where λ is the arrival rate and $1/\mu$ is the maximum over all possible service phase configurations (assuming all servers are busy) of the expected time until the next service completion.*

Proof: The proof follows that of Proposition 1 except that we consider the chain imbedded at the start (still denoted by η_t) of each service. Then Y_t, the expected number of arrivals before the next start of a service, is identically equal to zero if all servers are not busy at time η_t. If all servers are busy, Y_t depends on the server configuration $J \in \Sigma$. However, since μ is calculated assuming a worst case scenario (longest time to next service completion), we have $E(Y_t|J) \leq \rho$ for all $J \in \Sigma$ so that the mean drift satisfies

$$d_{(i,J)} \leq \rho - 1 + (\max_k\{p_k\})^i \to \rho - 1 \quad as \quad i \to \infty \qquad \square$$

Only relatively simple phase type distributions, with small dimension (one or two; maybe three) will be practical for multiserver models since the dimension of the state space increases exponentially as the number of phases increases. However, it may be desirable to obtain at least a stability condition for more complex systems. For the case of exponential servers, the stability condition is given by the usual $\lambda/s\mu < 1$ where μ is the service rate for each server. If a service phase j is such that the next phase of service is always k (i.e. all paths to service

completion from phase j pass through k) it is not necessary to consider phase k in the calculation of μ and we can ignore elements of Σ in which any server is in phase k since it will clearly not lead to the worst case. For example, if the service time distribution is a mixture of two erlangs, of order n and $n+1$ with identical rate parameter, then we only need to consider the case where all servers are in the phase furthest from service completion. Recall this distribution is good for fitting the first two moments of a distribution when the coefficient of variation is smaller than 1.0. Similarly, if the service time distribution is hyperexponential, we can assume that all servers are in the phase with the longest sojourn time. More complex distributions pose an interesting problem in combinatorial optimization but we will not pursue this here. Two dimensional phase type distributions are either hyperexponential (when the coefficient of variation is greater than 1) or a generalized Erlang (c.v. less than one). The former case we have already mentioned and, in the latter case, we can assume that all servers are in the first phase. The time until the next service can then be represented as a phase type distribution with tridiagonal subgenerator (keep track of the number of servers in phase 2) and μ can be calculated in $O(s)$ time.

4.3 APPROXIMATION METHOD

Although we could construct the generator for an M/PH/1 retrial queue with phase type retrial times and solve for the stationary distribution in principle, it would require an immense computational effort for any retrial time distributions with representations of dimension larger than about 2. In fact we we will use two dimensional representations (as in Yang et al) to obtain exact solutions in order to evaluate the performance of the approximation.

The key to the approximation is that interretrial times are gen-

erally shorter than service times so that most customers will execute a number of retrials during each service. Yang et al assume that the distribution of the elapsed retrial time for each customer after a service completion takes on its asymptotic form $m(x) = \int_0^x \theta(1 - T(u))du$ (see Ross, 1983) . If the retrial time has a phase type distribution, the obvious analog of elapsed retrial time is the phase of retrial and, if (α^T, T) represents the retrial time distribution, then, after a sufficiently long time, the distribution of phases is given by the solution π^T of the equation $\pi^T(T + T^0\beta^T) = 0$. If we assume that every customer has this average distribution π of phases after each service completion, the time until the next retrial would simply be the minimum of i retrial times (with i customers in the orbit) each of which has the distribution (π^T, T). Therefore the approximate distribution of the time until the next retrial has cumulative distribution

$$F(t) = 1 - [\pi^T exp(Tt)e]^i.$$

Unfortunately, although this is a phase type distribution, unless the retrial time distribution posesses some special structure, we can only gaurantee the existence of representations with dimensions n^i or i^n, which are far too large for large i. On the other hand, we can approximate this distribution with a phase type distribution of much smaller dimension. For each value of i, thwe construct an approximate phase type distribution (α_i^T, T_i) for the time until the next request from orbit by matching the moments of the exact distribution. If the coefficient of variation c of the distribution is smaller than 1, we match the first two moments to a mixture of two Erlang distributions of order n and $n+1$ ($n \leq c^2 \leq n+1$) with common rate parameter. If $c > 1$, we match the first three moments to a hyperexponential distribution of order 2 by aplying the method in Altiok (1985). We assume that, after each service completion, the time until the next retrial has the resulting distribution (α_i, T_i) when there are i customers in the orbit.

4.4 STATIONARY DISTRIBUTION

The approximate generator has the form

$$Q = \begin{bmatrix} A_{-1,1} & A_{-1,0} & 0 & 0 & 0 & \dots \\ A_{02} & A_{01} & A_{00} & 0 & 0 & \dots \\ 0 & A_{12} & A_{11} & A_0 & 0 & \dots \\ 0 & 0 & A_{22} & A_{21} & A_0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{4.1}$$

where:

$$A_{-1,1} = -\lambda \qquad A_{-1,0} = \lambda\beta^T$$

$$A_{02} = S^0 \qquad A_{01} = S - \lambda I \qquad A_{00} = [\,0 \quad \lambda I\,]$$

$$A_{12} = \begin{bmatrix} T_1^0\beta^T \\ 0 \end{bmatrix} \qquad A_0 = \lambda \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} \tag{4.2}$$

and for $i \geq 1$:

$$A_{i2} = \begin{bmatrix} 0 & T_i^0\beta^T \\ 0 & 0 \end{bmatrix} \qquad A_{i1} = \begin{bmatrix} T_i - \lambda I & \lambda e_i\beta^T \\ S^0\alpha_i^T & S - \lambda I \end{bmatrix}. \tag{4.3}$$

e_i represents a column vector of ones with the same dimension as the representation (α_i, T_i). The level i represents the number of customers in the orbit except that we have split the server idle state from the server busy states in the case of no customers in the orbit. This is done to avoid solving a boundary condition to obtain the stationary probabilities for the lowest level. The subdiagonal blocks have rank one ($A_{i2} = \gamma_i\beta_{i-1}^T$ where γ_i and β_{i-1} are column vectors) and so we can use Propositions 1 and 2 from Chapter 1 to obtain the stationary distribution x. We have $x = [x_0, x_1, \dots]$ and the x_i are obtained by the recusion

$$x_{i+1} = x_i R_i.$$

The rate matrices R_i are given by

$$R_i = -A_{i0}(A_{i+1,1} + A_{i+1,0}e\gamma^T)^{-1}$$

where $\gamma^T = [0 \quad \beta^T]$. Substituting from (4.2) and (4.3) yields

$$R_{-1} = \lambda\beta^T(\lambda(I - e\beta^T) - S)^{-1}$$

,

$$R_0 = \left[(1-p_1)e\alpha_1^T(\lambda I - T_1)^{-1}, \lambda[\lambda(e\beta^T - I) + p_1 S^0\beta^T - S]^{-1}\right]$$

$$R_i = \begin{bmatrix} 0 & 0 \\ (1-p_{i+1})e\alpha_{i+1}^T(\lambda I - T_{i+1})^{-1} & \lambda[\lambda(e\beta^T - I) + p_{i+1}S^0\beta^T - S]^{-1} \end{bmatrix}$$

where $p_i = \alpha_i(\lambda I - T_i)^{-1}e_i$ is the probability that an exponential interarrival time completes before the end of an interretrial time with distribution (α_i, T_i). Of course, in practice we must truncate and normalize in order to solve for the scalar x_0. If we let $x_i = [x_i^0, x_i^1]$, where the elements of x_i^0 and x_i^1 correspond to states in which the server is idle or busy respectively, then we have

$$x_i^0 e_i = (1-p_i)^2 x_{i-1}^1 e$$

$$x_i^1 = x_{i-1}^1 \lambda(\lambda(I - e\beta^T) - p_i S^0\beta^T - S)^{-1}.$$

Clearly, if we are interested only in the joint distribution of the number of customers in orbit and the number of customers in service, we require only the p_i for the approximate retrial distributions and we can omit the step of approximating them by phase type distributions. This is not surprising since that is exactly the information required for the approximation in Yang et al. (1994). However, we are interested in obtaining more information about the queue and so will require the forms (α_i, T_i) of the approximate retrial time distributions. This will also be required for more complex models such as the MAP/PH/1 retrial queue with phase type retrial times, although we do not examine these explicitly here.

4.5 NUMBER OF RETRIALS AND WAITING TIME

We can obtain the distribution of the number of retrials by tracking one customer through all of his retrials with the queue functioning independently in the background. We evolve the system from one retrial to the next until the server is found to be idle at the epoch of one of our customer's retrials. To do this, we must construct a generator which describes the evolution of both the customer and the queue. This generator has the form

$$Q' = \begin{bmatrix} Q_0 + Q_t & Q_s \\ 0 & Q \end{bmatrix}$$

where Q is the generator in (4.1) and Q_s and Q_t are block diagonal with blocks A_i^s and A_i^t respectively :

$$A_i^s = \begin{bmatrix} 0 & e_i\beta^T \otimes T^0 \\ 0 & 0 \end{bmatrix} \qquad A_i^t = \begin{bmatrix} 0 & 0 \\ 0 & I \otimes T^0\alpha^T \end{bmatrix}.$$

Q_s corresponds to transitions in which the tagged customer retries for service when the server is idle and begins service. Q_t corresponds to transitions in which the tagged customer finds the server busy upon retrying for service and begins another retrial period. Q_0 corresponds to the rest of the possible transitions and has the form

$$Q = \begin{bmatrix} A_{0,1}^0 & A_0^0 & 0 & \cdots & \\ A_{12}^0 & A_{11}^0 & A_0^0 & 0 & \cdots \\ 0 & \ddots & \ddots & \ddots & \\ \vdots & & \ddots & \ddots & A_0^0 \\ & & & A_{N2}^0 & A_{N1}^0 \end{bmatrix}$$

where:

$$A_{i1} = \begin{bmatrix} (T_i - \lambda I) \otimes I + I \otimes T & \lambda I \otimes e_i\beta^T \\ I \otimes S^0\alpha_i^T & (S - \lambda I) \otimes I + I \otimes T \end{bmatrix}$$

$$A^0_{N1} = \begin{bmatrix} (T_i - \lambda I) \otimes I + I \otimes T & \lambda I \otimes e_i \beta^T \\ I \otimes S^0 \alpha_i^T & S \otimes I + I \otimes T \end{bmatrix}$$

$$A^0_0 = \begin{bmatrix} 0 & 0 \\ 0 & \lambda I \end{bmatrix} \qquad A^0_{i2} = \begin{bmatrix} 0 & I \otimes T^0_i \beta^T \\ 0 & 0 \end{bmatrix}$$

Consider now the absorbing Markov chain with generator

$$Q'' = \begin{bmatrix} Q_0 & Q_s & Q_t \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Since arrivals follow a Poisson distribution, arriving customers see time averages. Thus when our customer arrives, the probability that he receives immediate service is just $P(n = 0) = \sum_i x^0_i e_i$, the stationary probability that the server is idle. The conditional distribution of states of the combined customer-queue system at his arrival given that he does not receive service immediately is given by $x' = [x'_0, x'_1, \ldots]$ where $x'_i = (1 - P(n = 0))^{-1}[(0, x^1_i) \otimes \alpha^T]$. The probability that he enters service at his first retrial (given he does not enter service immediately) is given by $-x' Q_0^{-1} Q_s e$. If he does not enter service at this point, the stationary distribution of states is given by $x' Q_0^{-1} Q_t / x' Q_0^{-1} Q_t e$. We can continue this process, going from one retrial to the next, to obtain the distribution of the number of retrials:

$$P(n = k | n > 0) = x'(-Q_0^{-1} Q_t)^{k-1}(-Q_0^{-1} Q_s)e.$$

We can use Algorithm 1.1 from Chapter One to multiply an arbitrary row vector by Q_0^{-1} and thus, eventually calculate the probabilities above. The conditional waiting time has a phase type distribution with representation $(x', Q_0 + Q_t)$ and we can use the methods of Section 1.8 to obtain this distribution via the randomization method. As in Section 1.8, we could eliminate the idle states above some level to obtain an approximation which is also a lower bound for this distribution however obtaining an upper bound would be significantly more complex

because of the non-exponential nature of the approximate interretrial time distributions (α_i, T_i). We might also approximate the waiting time distribution $p_w(t)$ by employing the distribution of the number of retrials n according to

$$p_w(t) \approx P(n=0)\delta(t) + \sum_{k=1}^{\infty} P(n=k) f^{(n)}(t)$$

where $f^{(n)}$ is the nth convolution of the retrial time distribution with itself. We did not attempt any of these approximations so we cannot comment on their effectiveness, however we did investigate the performance of the approximation for the first two moments of the waiting time. If we only require moments of the waiting time, they can be obtained by applying the relation

$$M_j = (-1)^j j! x'(Q_0 + Q_t)^{-j} e.$$

Algorithm 1.1 can also be used to operate on vectors with the tridiagonal subgenerator $(Q_0 + Q_t)^{-1}$ in order to evaluate the moments.

4.6 EXACT SOLUTION FOR PH_2 RETRIAL TIMES

If inter-retrial times have a phase type distribution with a two dimensional representation, we can obtain the stationary distribution and the moments of the waiting time in a reasonable amount of time without approximating the system. The state space is given by $\sigma = \{(i,j,k,\ell) | i = 0,1,\ldots; j = 0,1; k = 0,1,\ldots,i; \ell = 1,\ldots,nj\}$ where n is the dimension of the service time representation. i, j, k, represent respectively the number of customers in orbit, in service and in phase two of retrial. ℓ represents the current phase of service. The generator has the usual block tridiagonal form (1.1) with

$$A_{io} = \begin{bmatrix} 0 & 0 \\ 0\lambda U_i & \end{bmatrix} \qquad A_{i2} = \begin{bmatrix} 0 & D_i \otimes \beta^T \\ 0 & 0 \end{bmatrix}$$

$$A_{i1} = \begin{bmatrix} C_i - \lambda I & \lambda I \otimes \beta^T \\ I \otimes S^0 & (C_i + D_i U_{i-1} - \lambda I) \otimes I + I \otimes S \end{bmatrix}$$

where C_i, D_i and U_i are as defined in Section (1.10). In order to determine where to truncate the generator, we assume that when there are more than N customers in the orbit, the server begins serving the next customer immediately after each service completion. We assume that the retrial process ceases when the $N+1$st customer joins the orbit, and that the retrial process $(C_N, D_N U_{N-1})$ is restarted in its staionary distribution γ_N $(\gamma[C_N + D_N U_{N-1}] = 0)$ when a service completion leaves N customers in the orbit. The approximate generator has the form

$$Q = \begin{bmatrix} A_{00} & A_{01} & & & & & & \\ A_{12} & A_{11} & A_{10} & & & & & \\ & \ddots & \ddots & \ddots & & & & \\ & & A_{N2} & A_{N1} & A_{N0} & & & \\ & & & A_{N+1,2} & A_1 & A_0 & & \\ & & & & A_2 & A_1 & A_0 & \\ & & & & & \ddots & \ddots & \ddots \end{bmatrix}$$

where $A_0 = \lambda I$, $A_1 = S - \lambda I$, $A_2 = S^0 \beta^T$, $A_{N+1,2} = [0 \quad \gamma_N \otimes S^0 \beta^T]$ and

$$A_{N0} = \begin{bmatrix} 0 \\ \lambda e \otimes I \end{bmatrix}.$$

We gradually increase the value of N until the probability (in the approximate queue) of the orbit containing more than N customers is less than some small number ϵ. We used $\epsilon = 10^{-5}$ for the numerical examples we considered. Once N is determined, we truncate the orbit above N so that all customers who arrive when there are N in the orbit are lost to the system forever.

In order to obtain the distribution or moments of the waiting time, we must track a tagged customer from her arrival until she begins service. The probability of immediate service is given by $x(I - P_b)e$ where

x is the stationary distribution and P_b is the projection matrix corresponding to states in which the server is busy. The waiting time for customers who do not receive immediate service has a phase type distribution with representation (x', Q') where $x' = (xP_b \otimes \alpha)/xP_b e$ (recall (α, T) is the retrial time representation) and Q' is a block tridiagonal generator with blocks $A'_{i0} = A_{i,0} \otimes I$, $A'_{i,2} = A_{i,2} \otimes I$ and

$$A'_{i1} = A_{i1} \oplus T + \begin{bmatrix} 0 & 0 \\ 0 & I \otimes T^0 \alpha^T \end{bmatrix}.$$

We can obtain the distribution of the waiting time by applying the randomization method as in Section 1.8 although the computation will be extremely time consuming unless the utilization of the queue is low. We could also find the distriution of the number of retrials by a method analogous to the one applied in Section 4.5. These methods are computationally intensive although not difficult to implement. In the following section, we obtain the moments of the waiting time for a number of different examples. Since there does not appear to be any particular structure we can exploit to speed up computation of the stationary distribution and the waiting time moments (which require multiplying vectors by powers of Q^{-1}), we use Algorithm 1.1 directly when applying block Gaussian elimination. For some of the numerical examples considered with high traffic intensities ($\rho = \lambda/\mu$) or low retrial rates (θ), the computational time required to implement the block Gaussian elimination for the exact model was excessive. In these cases, we still used block Gaussian elimination to solve the approximate model, because the method provides a good scheme for deciding where to truncate the system, but we used block Gauss-Seidel iteration to obtain the solutions for the exact model. The block Gauss-Seidel method is efficient here because the diagonal blocks of the generator are block tridiagonal if we re-order the state space, grouping together the states in each level with the same number of customers in phase 2 of retrial. The effect of multiplying a vector by the inverse of one of the diagonal blocks A_{i1} can then

be calculated in $O(i)$ time via block Gaussian elimination once the reduction phase has been performed on the blocks of A_{i1}. Since we need to perform the latter operation only once and the former operation many times in the block Gauss-Seidel method, this affords a significant savings in computational time if we are willing to accept the corresponding loss of accuracy that results from using an iterative scheme.

4.7 NUMERICAL RESULTS AND CONCLUSIONS

We demonstrate the performance of the approximation in predicting the first two moments of the waiting time for a number of different examples. Although we could estimate the first moment from the stationary distribution by applying Little's law, we calculate it in both the approximate and exact models via the general formula for moments. Since we need to apply the inverse of the subgenerator twice anyway in order to obtain the second moment, this does not require any extra computational effort and it allows us to verify Little's law explicitly as a consistency check.

We tested our approximation on the set of examples used by Yang et al. : M/M/1, M/E_2/1 and M/H_2/1 retrial queues with two stage Erlang or two stage hyperexponential retrial time distributions. The Erlang service time distributions had a mean service time of 1.0 and the hyperexponential service times had the representation (β, S) with $\beta = (.1, .9)$ and

$$S = \begin{bmatrix} -.22 & 0 \\ 0 & -1.73 \end{bmatrix}.$$

The Erlang retrial time distributions were taken to have mean $1/\theta$ and the hyperexponential retrial times were assumed to have the form (α, T) with

$$\alpha = (\frac{1}{K+1}, \frac{K}{K+1}) \qquad T = \begin{bmatrix} -\frac{2\theta}{K+1} & 0 \\ 0 & -\frac{2K\theta}{K+1} \end{bmatrix}$$

where $K = c^2 + \sqrt{c^4 - 1}$ and $c > 1$ is the coefficient of variation. Tables 4.1, 4.2 and 4.3 present the values of the mean and coefficient of variation of waiting time for the exact models and the relative error in these quantities (aproximate/exact-1) for the approximate models of the M/M/1, M/E_2/1 and M/H_2/1 (with $c = 1.5$) retrial queues respectively. Table 4.4 presents the same quantities for the M/M/1 retrial queue for various values of the retrial time coefficient of variation c.

The relative error in both moments seems to decrease with the retrial rate θ for queues with Erlang retrial times and increase with the retrial rate for queues with hyperexponential retrial times. Also the approximation appears to perform better when the coefficient of variation of the service time and the retrial time is closer to 1.0. The approximation becomes poorer as the utilization ρ grows larger but the effect is minimal if the coefficients of variation for the service and retrial times are close to 1.0. In general, the approximation performs quite well as long as we do not stray too far from the cv of the exponential distribution for service and retrial times. Most of the relative errors are close to 1%. The approximation performs well at low utilizations for all service and retrial time distributions considered.

Table 4.1: Waiting time for $M/E_2/1$ retrial queue

		E_2 retrial times				H_2 retrial times			
ρ	θ	mean	% error	cv	% error	mean	% error	cv	% error
	1	.1737	.17	1.920	-.00	.2158	.44	2.093	-.01
.1	3.3	.1094	.09	1.972	.00	.1295	1.0	2.006	.10
	10	.0918	.04	2.021	.00	.1000	1.5	2.007	.23
	1	.6778	.41	1.471	-.01	.8209	1.9	1.568	.02
.3	3.3	.4240	.23	1.518	-.01	.4928	3.7	1.530	.39
	10	.3546	.08	1.554	-.01	.3832	5.3	1.542	.69
	1	2.4310	.45	1.265	-.04	2.791	5.4	1.301	.42
.6	3.3	1.5000	.37	1.320	-.03	1.680	9.9	1.314	1.03
	10	1.2460	.17	1.352	-.02	1.323	13	1.300	1.23

Table 4.2: Waiting time for $M/H_2/1$ retrial queue

		E_2 retrial times				H_2 retrial times			
ρ	θ	mean	% error	cv	% error	mean	% error	cv	% error
	1	.3597	.05	2.195	-.02	.4114	1.82	2.187	.14
.1	3.3	.2961	.02	2.319	-.01	.3168	1.60	2.282	.15
	10	.2783	.01	2.365	-.00	.2862	1.96	2.345	.08
	1	1.399	.04	1.661	-.05	1.563	4.82	1.653	.39
.3	3.3	1.146	.02	1.749	-.03	1.211	5.47	1.725	1.43
	10	1.075	.00	1.782	-.01	1.100	6.33	1.769	.23
	1	4.959	-.02	1.395	-.04	5.342	10.7	1.389	.97
.6	3.3	4.030	.01	1.460	-.01	4.184	13.8	1.448	.76
	10	3.768	.00	1.484	-.01	3.828	15.1	1.478	.53

Table 4.3: Waiting time for M/M/1 retrial queue

		E_2 retrial times				H_2 retrial times			
ρ	θ	mean	% error	cv	% error	mean	% error	cv	% error
	1	.2009	.14	1.948	-.00	.2471	1.2	2.078	.01
.1	3.3	.1372	.04	2.029	-.01	.1580	1.1	2.034	.15
	10	.1196	-.01	2.082	-.02	.1278	1.8	2.062	.22
	1	.7832	.31	1.493	-.03	.9390	3.2	1.564	.13
.3	3.3	.5314	.18	1.556	-.01	.6019	4.0	1.552	.55
	10	.4619	.07	1.593	-.00	.4901	6.0	1.579	.66
	1	2.802	.28	1.285	-.06	3.191	6.0	1.308	.64
.6	3.3	1.878	.21	1.344	-.04	2.059	11	1.333	1.2
	10	1.622	.09	1.374	-.02	1.696	15	1.363	1.1

Table 4.4: Waiting time for M/M/1 retrial queue

cv of retrial times		.707		2		3		4	
ρ	θ	cv	% error	cv	% error	cv	% error	cv	% error
	.1	1.457	-.04	1.832	-.05	2.142	-.05	2.426	-.04
.3	1	1.493	-.03	1.646	-.08	1.836	-.31	2.029	-.36
	10	1.593	-.00	1.575	.88	1.585	.68	1.621	.25
	.1	1.203	-.05	1.418	-.08	1.619	-.17	1.813	-.19
.6	1	1.285	-.06	1.347	.40	1.453	-.31	1.572	-.71
	10	1.374	-.02	1.358	1.5	1.356	1.6	1.366	1.2
.	.5	1.178	-.08	1.235	.64	1.318	-.18	1.413	-.89
.8	1	1.222	-.07	1.244	1.4	1.297	.56	1.368	-.49
	10	1.314	-.02	1.305	1.2	1.300	1.6	1.300	1.7

CONCLUSION

We have applied matrix analytical methods to a number of key models of retrial queues. We have found that this allows the inclusion of significantly more general arrival and service processes than those appearing in other models. The use of the Markovian arrival process and the phase type distribution allows for the straightforward construction of numerically tractable problems without sacrificing too much freedom in specifying the arrival and service processes. Errors introduced by numerical approximations have been bounded analytically where possible and demonstrated empirically to be small for a significant set of models where analytical bounds were not available.

The transparent probabilistic interpretation of the calculations which are performed should make it easy to extend the method to obtain information about other performance criteria which may not have been considered here. The goal of this thesis has been the development of some general methods for the analysis of retrial queues. There are no doubt many retrial queues not considered here which can benefit from these methods. Also, a greater understanding of the models included may be gained by applying the methods to more extensive numerical experiments. Since our focus was on the development of the methods, this is somewhat outside the scope of the thesis; however, the tools are now available for this future research.

REFERENCES

T. Altiok, On the phase-type approximations of general distributions. *IIE Transactions,* **17** , 1985, 110-116.

J.R. Artalejo, Explicit formulae for the charicteristics of the M/H_2/1 retrial queue. *Queueing Systems,* **44** , 1993, 309-313.

S. Asmussen and O. Nerman, Fitting phase-type distributions via the EM algorithm. *Symposium i Anvendt Statistik* , Copenhagen, Jan. 21-23, 1991, 335-346.

G.P. Basharin, and V.A. Naumov, Losungsmethode fur lineare algebraische gleichungssysteme stationarer charakteristiken, *Handbuch der Bedienungstheorie* , **1** , Berlin, Akademie-Verlag, 1983.

R. Bellman, *Introduction to Matrix Analysis.* McGraw Hill, Toronto, 1970.

G. Birkhoff and S. Maclane, *A Survey of Modern Algebra* , The MacMillan Company, New York, 1953.

A. Bobbio, A. Cumani, A. Premoli, and O. Saracco, Modelling and identification of non-exponential distributions by homogeneous Markov processes. *NCSR R23* , **1** , 1980, 373-392.

R.K. Boel and N.W. Talat, Performance analysis and optimal threshold policies for queueing systems with several heterogeneous servers and markov modulated arrivals, preprint, 1994.

A Brandt and G. Last, On the pathwise comparison of jump processes driven by stochastic intensities, *preprint, Humboldt-Universitat zu*

Berlin , 1993.

L. Bright and P.G. Taylor, Calculating the equilibrium distribution in level dependent quasi-birth-and-death processes, *Stochastic Models,* **11** , No. 3, 1995.

J.L. Carrol, A. Van De Liefvoort, and L. Lipsky, Solutions of M/G/1//N-type Loops with extensions to M/G/1 and GI/M/1 queues. *Operations Research* ,**30** , 1982, 490-514.

S. Chakravarthy, *Private Communications,*March 1993

B.D. Choi, An M/M/1 retrial queue with control policy and general retrial times, *Queueing Systems Theory and Applications,* **14** , 275-292, 1993.

E.J. Coyle and B. Liu, A matrix representation of CSMA/CD networks. *IEEE Transactions on Communications* , **COMM-33** , 1985, 53-64.

A.G. De Kok, Algorithmic methods for single server systems with repeated attempts. *Statistica Neerlandica* , **38** , 1984, 23-32.

J.E. Diamond and A.S. Alfa, Matrix analytical methods for M/PH/1 retrial queues. *Stochastic Models, 11, No.3* , 1995.

A. Elldin, Approach to the theoretical description of repeated call attempts. *Ericsson Technics* , **23** , (1967), 346-407.

G.I. Falin, A single line system with secondary orders, *Eng. Cybernet. Rev.* , **17** , 1979, 76-83.

G.I. Falin, An M/M/1 queue with repeated calls in the presence of

persistence function. Paper #1606-80, *All Union Institute for Scientific and Technical Information.*Moscow, 1980.

G.I. Falin, Calculation of probabilistic characteristics of a multi-channel queue with reapeated calls, *Vestnik Mosk. Univ. Ser 15, Vychisl. Mat. Kibernet* , **3** , 1983, 66-69.

G.I. Falin, On sufficient conditios for ergodicity of multichannel queueing systems with repeated calls. *Adv. Appl. Prob.*, **16** , 1984, 447-448.

G. Falin, Single line repeated orders queueing systems. *Optimization* , **17** ,(1986), 649-667.

G. Falin, A survey on retrial queues. *Queueing Systems* , **7** , 1990, 127-168.

G. Falin, On the virtual waiting time in an M/G/1 retrial queue. *J. Appl. Prob.,* **28** , 1991, 446-460.

D.P. Gaver, P.A. Jacobs and G. Latouche, Finite birth-and-death models in randomly changing environments, *Adv. App. Prob.* , **16** , 1984, 715-731.

F. Gillent, and G. Latouche, Semi-explicit solutions for M/PH/1-like queueing systems. *EJOR,* **13** , 1983, 151-160.

B.S. Greenberg, Queuing systems with returning customers and the optimal order of tandem queues, Ph. D. thesis, University of California, Berkely,1986.

B.S. Greenberg, and R.W. Wolff, An upper bound on the performance of queues with returning customers. *J. Appl. Prob.* , **24** , 1987,

466-475.

B.S Greenberg, M/G/1 queueing systems with returning customers. *J. Appl. Prob.* , **26** , 1989, 152-163.

O. Hägström, S. Asmussen and O. Nerman, EMPHT - A program for fitting phase type distributions, *Studies in Statistical Quality , Control and Reliability,* Technical report 1992:4, Mathematical Statistics, Chalmers University of Technology, The University of Göteborg, Sweden, 1992.

B. Hajek, Birth-death processes on integers with phases and general boundaries, *J. App. Prob.* , **19** , 1982, 488-499

I.I. Homitchkov, Generating functions of state probabilities of a single line queue with repeated calls, *Vestnik Beloruss Univ. Ser. 1* **1** , 1987, 51-55, (in Russian).

I.I. Homitchkov, Single line queue with repeated calls and Cox innput process of second order, *Vestnik Beloruss Univ.* **1** , 1988, 70-71, (in Russian).

M.A. Johnson, and M.R. Taaffe, Matching moments to phase distributions: Nonlinear programming approaches. *Research Memorandum No. 8814* , School of Industrial Engineering, Purdue University, 1988.

G.I. Jonin and J.J. Sedol, Full available groups with repeated calls and time of advanced service. *Proc. 7th Int. Teletrafffic Congress* , Stockholm (1973), 137/1-137/4.

V.A. Kapyrin, A study of the stationary characteristics of a queueing system with recurring demands. *Cybernetics* , **13** , (1977),584-590.

J. Keilson, J. Cozzolino and H. Young, A service system with unfilled requests repeated. *Operations Research* , **16** , 1968, 1126-1

L. Kleinrock, *Queueing Systems, Volume 1: Theory.* John Wiley & Sons, New York, 1975.

Y.N. Kornishov, Waiting positions for overloading trunks. *Elektrosvyaz* , **7** , (1974), 32-39.

H.M. Liang, *Retrial queues (queueing system, stability condition, K-ordering)* , Ph. D. thesis, University of North Carolina at Chapel Hill, 1991.

H.M. Liang and V.G. Kulkarni, Monotonicity properties of single-server retrial qu eues, *Stochastic Models* , **9** , 373-400,1993.

D.M. Lucantoni, New results on the single server queue with a batch Markovian arrival process, *Stochastic Models* , **7** ,1-46, 1991.

M. Marcus,and H. Minc *A survey of Matrix theory and Matrix inequalities.* Allyn and Bacon, Boston, 1964.

R.A. Marie and J.M. Pellaumail, Steady state probabilities for a queue with a general service discipline and state dependent arrivals. *IEEE Trans. Soft. Eng.* , **SE-9** , 1983, 109-113.

W.A. Massey, Stochastic orderings for Markov processes on partially ordered spaces, *Mathematics of Operations Research* , **12** , 1987, 350-367.

V.A. Naumov, *Systems of equilibrium equations* ,
Moscow: Peoples Friendship University Press, 1985 (in Russian).

M.F. Neuts, *Matrix-geometric solutions in stochastic models: an algorithmic approach.* Johns Hopkins University Press, Baltimore, 1981.

M.F. Neuts, and M.F. Ramalhoto, A service model in which the server is required to search for customers. *J. Appl. Prob.* , **21** , 1984, 157-166.

M.F. Neuts and B.M. Rao, Numerical investigation of a multiserver retrial model, *Queueing Systems* , **7** , 1990, 169-190.

C.E.M. Pearce, On some basic properties of the inhomogeneous quasi-birth- and-death process. *Preprint* , Dept. Applied Mathematics, University of Adelaide, (1994).

B. Pourbabai, Analysis of a G/M/K/O queueing loss system with heterogeneous servers and retrials. *International Journal of Systems Science* , **18** , (1987),985-992.

V. Ramaswami and G. Latouche, A general class of markov processes with explicit matrix-geometric solutions. *OR Spektrum* ,**8** , 1986, 209-218.

G.E. Rideout, A study of retrial queueing systems with buffers. *M.A.Sc¿ thesis, Dept. of Industrial Engineering, University of Toronto* , (1984).

S.M. Ross, *Stochastic Processes* , Wiley New York, (1983).

D. Sonderman and B. Pourbabai, Single server stochastic recirculation systems, *Computers and Operations Research* , **14** , 1987, 75-84.

P.M. Snyder and W.J. Stewart, Explicit and iterative numerical ap-

proaches to solving queueing models. *Operations Research* **33** , 1985, 183-202.

S. Stepanov, Optimal calculation of characteristics of models with repeated calls, *Proc. 12th Int. Teletraffic Congress* , Torino, 1988.

R.S. Varga, *Matrix Iterative Analysis.* Prentice-Hall, New Jersey, 1962.

W. Whitt, On approximations for queues III: Mixtures of exponential distributions. *BSTJ* , **63** ,1984.

R.I. Wilkinson and R. Radnik, Theories for toll traffic engineering in the USA. *B.S.T. J.* , **35** , (1956), 421-507.

T. Yang, *A broad class of retrial queues and the associated generalized recursive technique* , Ph. D. Thesis , University of Toronto, 1990.

T. Yang, M.J.M. Posner and J.G.C. Templeton, C//a/M/s/m retrial queue: a computational approach, *ORSA Jounal on Computing* , **4** , 182-191, 1992.

T. Yang and J.G.C. Templeton, A survey of retrial queues. *Queueing Systems,* **2** ,1987, 201-233.

T. Yang, M.J.M. Posner, J.G.C. Templeton and H. Li, An approximation method for the M/G/1 retrial queue with general retrial times. *EJOR* **76** , (1994), 552-562.

J. Ye and S.Q. Li, Analysis of multimedia traffic queues with finite buffer and overload control - Part I: Algorithm, *Proc. of IEEE Infocom '91* , April, 1991, 1464-1474.