1

MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS
STANDARD REFERENCE MATERIAL 1010a
(ANSI and ISO TEST CHART No. 2)

National Library
of Canada

Bibliothèque nationale
du Canada

Canadian Theses Service

Service des thèses canadiennes

Ottawa, Canada
K1A 0N4

NOTICE

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

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

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

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

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

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

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

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

Applications of Matrix- Analytic Solutions to Queueing Systems Arising in Integrated Services Digital Network

by

Benjamin Bin-yong NGO, B. Eng. (Electrical)

**A thesis submitted to the
Faculty of Graduate Studies and Research
in partial fulfillment of
the requirements for the degree of
Master of Engineering**

**Ottawa-Carleton Institute for Electrical Engineering
Department of Systems and Computer Engineering
Carleton University
Ottawa, Ontario
Canada**

August 9, 1990

© Benjamin Bin-yong NGO, 1990

NOTICE

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

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

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

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

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

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

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

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

ISBN 0-315-60455-7

The undersigned hereby recommend to
the Faculty of Graduate Studies and Research
acceptance of the thesis,

**“Applications of Matrix-Analytic
Solutions to Queueing Systems Arising
in Integrated Services Digital Network”**

submitted by

Benjamin Bin-yong NGO
B. Eng. (Electrical)

in partial fulfillment of the requirements
for the degree of Master of Engineering

Thesis Supervisor

Chairman, Department of Systems and Computer Engineering

Carleton University

September 5, 1990

Abstract

The matrix-analytic method has potential in solving queueing systems which arise in the ISDN environment. Classical methods such as the generating function approach are usually difficult to apply due to dimensionality constraint and numerical instability. Three classes of queueing system are analyzed via the matrix-analytic method, which include : 1) Selective-repeat ARQ scheme where previous analysis has limited applicability due to dimensionality problem; matrix-analytic method helps in relaxing the constraint in the exact analysis, and an approximation is derived to further extend the limit; 2) Access Control Strategies where it is generalized to allow both types of prioritized "service requests" to be queued and flexible multiple seizures of servers for one of the service types; strategies using non-preemptive priority and preemptive priority schemes are analyzed and compared; and 3) Correlated source arrival modelling and approximate analysis for the queueing behaviour of the low priority fixed length packets under the influence of correlated high priority fixed length packets at the packet transmission level of the multiplexer. Matrix-analytic method is found to be useful and numerically stable in obtaining probability distribution (which is often difficult with classical methods) wherever the queueing system can be modelled as "Markov chain/process of M/G/1 Type" or "Markov chain/process of GI/M/1 Type".

Acknowledgments

I would like to thank my thesis supervisor Dr. Hyong W. Lee for his guidance and encouragement during the course of this research. With his knowledge and enthusiasm in this area of research, his constant support, and numerous valuable discussion periods, a novice like me has gradually learned to discover the beauty and the complexity of queueing theory, and to appreciate its application.

I would also like to express my gratitude to my beloved family and Ms. Ida Leung, and my good friend Mr. Tak K. Tang for their support and encouragement during the two years of graduate studies and the preparation of this thesis.

I am grateful to Carleton University for the financial support that made my graduate studies possible.

Above all, I thank God for His abundant blessings and the opportunity for me to pursue further in knowledge.

Table of Contents

Abstract	iii
Acknowledgments	iv
List of Figures	viii
List of Acronyms	xi
Chapter 1 Introduction	1
1.1 Background	1
1.2 Objectives	3
1.3 Highlights of Thesis	4
1.4 Outline	5
Chapter 2 Matrix-Analytic Methodology	7
2.1 Introduction	7
2.2 Infinite Markov Chain of M/G/1 Type	8
2.3 Infinite Markov Chain of GI/M/1 Type	9
2.3.1 Quasi-Birth-and-Death (QBD) Process	10
2.3.2 Markov Process of Quasi-GI/M/1 Type with Complex Boundary behaviour (QGM1-CB process)	12
Chapter 3 Queueing Analysis of the Selective Repeat ARQ Scheme	13
3.1 Introduction	13
3.2 The Model	15
3.2.1 Basic system and the SR-ARQ protocol	15
3.2.2 Assumptions for the model	15
3.3 Analysis	17
3.3.1 Matrix-analytic approach (an exact analysis)	17
3.3.2 Approximate model and its analysis	19
3.4 Numerical Examples for SR-ARQ scheme	22
3.5 Extension of the Approximate Model	23
3.5.1 Channel with correlated block errors	23
3.5.2 System with FEC/SR-ARQ schemes	28
3.5.3 Numerical Examples	29
3.6 Conclusions	30

Chapter 4 Queueing Analysis of Traffic Access Control Strategies in Wide-Band Integrated Networks	33
4.1 Introduction	33
4.1.1 Channel Structure	34
4.1.2 Integrated Access Control Problem and Related Works	34
4.1.3 Objectives and Outlines	37
4.2 The Model	38
4.3 Analysis 1 — Non-Preemptive Priority M/M/c Model : Type 1–SR has priority over Type 2–SR	39
4.4 Analysis 2 — Non-Preemptive Priority M/M/c Model : Type 2–SR has priority over Type 1–SR	54
4.5 Analysis 3 — Preemptive Priority with Boundary M/M/c Model	60
4.6 Numerical Examples and Discussions	65
4.7 Conclusions	77
Chapter 5 A Study on the Markov Modulated Characterization of Packetized Videophone Traffic and related Statistical Multiplexer Performance with the application of Access Control Strategy	79
5.1 Introduction	79
5.2 The Videophone Source Models	82
5.2.1 Model A : Discrete-State Continuous-Time Markov Process	82
5.2.2 Model B : Markov Modulated Poisson Process	84
5.3 Queueing Performance of Statistical Multiplexer	87
5.3.1 Analysis — MMPP/D/c Queueing System for High Priority Packets	88
5.3.2 Analysis — Approximation of Queueing Performance for Low Priority Packets	92
5.3.2.1 Approximation 1	92
5.3.2.2 Approximation 2	96
5.4 Numerical Examples and Discussion	96
5.4.1 MMPP vs. DS-CTMP	96
5.4.2 Approximation of Low Priority (Data) Packet Queueing Performance	103
5.5 Conclusions	108
Chapter 6 Conclusions	111
6.1 Conclusions	111
6.2 Suggestions For Future Work	114
Appendix A Matrix Analytic Solution for M/G/1 type Paradigm	117

Appendix B Elements of transition probability matrix P for System with Correlated Block Errors	123
Appendix C Elements of the transition probability matrix P for System using FEC/CR-ARQ Scheme	125
Appendix D Algorithmic Solutions For Non-Preemptive Priority M/M/c Model (Analysis 1 — Sect. 4.3)	128
Appendix E Algorithmic Solutions For Non-Preemptive Priority M/M/c Model (Analysis 2 — Sect. 4.4)	143
Appendix F Algorithmic Solutions For Preemptive Priority M/M/c Model (Analysis 3 — Sect. 4.5)	158
Appendix G Procedures for Approximating Videophone Sources given as Minisources with MMPP	165
Appendix H Matrix Analytic Solution for MMPP/D/c Queueing System . .	171
Appendix I Analysis of Queueing Performance for Low Priority Packets — Approximation 2 (ref. Chap. 5 — Sect. 5.3.2)	178
References	182

List of Figures

Figure 3.1	Basic system for the SR-ARQ scheme	15
Figure 3.2	Queueing model for SR-ARQ scheme	16
Figure 3.3	Transmit queue length distribution ($N = 5, \alpha = 0.01$)	24
Figure 3.4	Transmit queue length distribution ($N = 5, \alpha = 0.1$)	25
Figure 3.5	Transmit queue length distribution ($N = 10, \alpha = 0.3$)	26
Figure 3.6	Transmit queue length distribution ($N = 100, \alpha = 0.3$)	27
Figure 3.7	Channel model with correlated block error rates	28
Figure 3.8	Transmit queue length distribution for the system with correlated block errors and the system using hybrid FEC/SR-ARQ scheme ($N=10$) . . .	31
Figure 4.1	A simplified view of the structure of ISDN access	34
Figure 4.2	Control of integrated access at the ISDN entrance node	35
Figure 4.3	The basic M/M/c two priorities queueing system	40
Figure 4.4	The partitioned state space (blocks) and transition scheme for Analysis 1	43
Figure 4.5	The partitioned state space (super-blocks) and transition scheme for Analysis 1	46
Figure 4.6	The partitioned state space (blocks) and transition scheme for Analysis 2	55
Figure 4.7	The partitioned state space (super-blocks) and transition scheme for Analysis 2	56
Figure 4.8	The preemptive priority with boundary M/M/c queueing system	60

Figure 4.9	The state space and transition scheme for PP system (Analysis 3)	62
Figure 4.10	Queue length distribution for (a) Type 1–SR and (b) Type 2–SR under various utilizations with Type 1–SR has non-preemptive priority over Type 2–SR ($b=6, c=20, \lambda_1=1.0, \mu_1=1.0, \mu_2=0.1$)	67
Figure 4.11	Queue length distribution for (a) Type 1–SR and (b) Type 2–SR under various utilizations with Type 2–SR has non-preemptive priority over Type 1–SR ($b=4, c=10, \lambda_1=0.1, \mu_1=0.1, \mu_2=1.0$)	68
Figure 4.12	Mean waiting time for (a) Type 1–SR and (b) Type 2–SR vs. system utilization (ρ) with system employs PP policy ($b=3, c=20, \lambda_1=0.1, \mu_1=0.1, \mu_2=1.0$)	70
Figure 4.13	Mean waiting time for (a) Type 1–SR and (b) Type 2–SR vs. system utilization (ρ) with system employs PP policy ($b=6, c=20, \lambda_1=1.0, \mu_1=1.0, \mu_2=0.1$)	71
Figure 4.14	Effect of NB group (c_2) on mean waiting time for (a) Type 1–SR and (b) Type 2–SR with PP policy ($b=4, c=20, \lambda_1/\lambda_2=10, \mu_1=1.0, \mu_2=0.1$)	72
Figure 4.15	Comparison of mean waiting time for (a) Type 1–SR and (b) Type 2–SR under various utilizations between NPP and PP policy ($b=3, c=10, \lambda_1=1.0, \mu_1=1.0, \mu_2=1.0$)	74
Figure 4.16	Comparison of queue length distribution for (a) Type 1–SR and (b) Type 2–SR between NPP and PP policy ($b=6, c=20, c_2=0, \lambda_1=1.0, \mu_1=1.0, \lambda_2=0.8, \mu_2=0.1, \rho=0.7$)	75
Figure 4.17	Comparison of queue length distribution for (a) Type 1–SR and (b) Type 2–SR between NPP and PP policy ($b=4, c=10, c_2=c, \lambda_1=0.1, \mu_1=0.1, \lambda_2=3.0, \mu_2=1.0, \rho=0.7$)	76
Figure 5.1	Videophone Source	80

Figure 5.2	Approximation for Videophone source output rate	80
Figure 5.3	State transition diagram for videophone source — Model A	84
Figure 5.4	Minisource model — Videophone source	85
Figure 5.5	The 2-State MMPP model	85
Figure 5.6	Approximation of aggregate videophone sources with MMPP	86
Figure 5.7	Queueing model for statistical multiplexer at packet transmission level	88
Figure 5.8	Index Of Dispersion curve of single videophone source with stat. matching at different time points	99
Figure 5.9	Queue length distribution of single videophone source with stat. matching at different time points (No. of servers in system $[c] = 1$)	100
Figure 5.10	Comparison of average queue size for videophone packets under different utilization (No. of servers $[c] = 1$)	101
Figure 5.11	Comparison of queue length distribution for videophone packets (No. of servers $[c] = 1$)	102
Figure 5.12	Comparisons of Approximations for queue length distribution for data packets (Utilization for videophone sources = 0.5, $c = 3$, $\bar{q} = 4$) . . .	105
Figure 5.13	Effects of $\bar{q}$ (queue length information) on the accuracy of approximation (Utilization for videophone sources = 0.5, $c = 3$)	106
Figure 5.14	Effects of $\bar{q}$ (queue length information) on the accuracy of approximation (Utilization for videophone sources = 0.65, $c = 3$)	107

List of Acronyms

ACK	- Positive Acknowledgement message in SR-ARQ scheme
ARQ	- Automatic Repeat Request
BBU	- Basic Bandwidth Unit
DS-CTMP	- Discrete State - Continuous Time Markov Process
FEC	- Forward Error Correction
GI/M/1	- single server queueing system with General Independent arrival process and Exponential service time distribution
ISDN	- Integrated Services Digital Network
ISDSN	- Integrated Services Digital Satellite Network
M/G/1	- single server queueing system with Poisson arrival process and General service time distribution
MMPP	- Markov Modulated Poisson Process
MMPP/D/c	- c-servers queueing system with Deterministic service time and arrival process as Markov Modulated Poisson Process
Mbps	- Mega-bits per second
msec.	- milliseconds
NACK	- Negative Acknowledgement message in SR-ARQ scheme
NB	- Narrow Band
NPP	- Non-Preemptive Priority
PP	- Preemptive Priority
QBD	- Quasi-Birth-and-Death Process
QGM1-CB	- Markov Process of Quasi-GI/M/1 Type with Complex Boundary behaviour
SR	- Service-Request
SR-ARQ	- Selective-Repeat Automatic Repeat Request Scheme
TDM	- Time Division Multiplexing
VMPP	- Versatile Markovian Point Process
WB	- Wide Band

Chapter 1

Introduction

1.1 Background

Integrated Services Digital Network (ISDN) refers to the integration of various types of communications services, such as video, data, and voice. These services are to be transported over a single digital facility at the user-network interface. It is foreseen that the advantages of having such an integrated transport include the flexibility to support a spectrum of service types and mixes "on demand", as well as the gain in efficiency by statistically multiplexing bursty traffic generated by several sources and service types. The bit rates among the various types of services may vary in the order of magnitudes, and obviously, traffic arrival patterns are not homogeneous and sometimes highly correlated. In order to support the various types of services with fairness and efficiency, proper policing of the ISDN resources' access by the services is required. Anticipation of the low error rate provided by the optical fiber link and increases in transmitting speed also make the conventional link level error control protocol inappropriate. In the realm of ISDN, error control is expected to be performed near the user-end and thus feedback time of the various types of control messages will considerably be longer than its transmission time. The realization of ISDN as the

common communications framework for the future still faces challenges. The analysis on the queueing system which arises in the ISDN environment is also a challenging task.

Classical analytic methods such as the complex variable analysis, generating function approach and Laplace Stieltjes transforms etc. have long been used for analyzing system performance. They are often difficult and risky in the numerical implementation. Although obtaining the mean or variance (e.g. the mean system time, the variance of queue length of the queueing system etc.) may not be very difficult with these methods, the probability distribution (from which all other forms of measures, e.g. variance, mean, third moments and percentiles can be easily obtained) is sometimes complex and next to impossible to obtain. In addition, small increases in the complexity of the physical description of the model often results in vast increases in dimensionality from which classical analytic methods' applicability is dramatically limited.

In late 70's, M.F. Neuts and his colleagues have proposed Matrix-analytic method [1, 2] to solve certain type of stochastic models which arise frequently in the computer and communications network environment. The method is based on the unifying structural properties of Markov Chain and pure probabilistic arguments, which results in highly stable numerical procedures. This method is a combination of iterative matrix methods and classical systems of differential equations. It can be used to solve the queue length distribution (and other measures) of certain stochastic model having an infinite-state embedded Markov chain with the stochastic matrix having a block partitioned structure. The beauty of this methodology lies in the reducing of the problem of dimensionality by decomposing the state space of the model into levels of blocks and solve each level at a time. The steady state probability vector for one level is interrelated with other levels through some special matrices, and from which the steady state probability distribution for the whole system can be found easily.

1.2 Objectives

The objectives of the thesis are to analyze the performance of three different classes of queueing systems, which arise in the ISDN environment, via the matrix-analytic method and verify the analysis via simulation wherever applicable. All the three queueing systems can be modelled to have structured embedded Markov chain (or process). The three queueing systems include :

1. Approximation of the transmit queue length distribution of a communications system employing selective-repeat automatic repeat request (SR-ARQ) scheme with rather long acknowledgment delay, which previous analysis [3, 4] has shown to be failed. Application of the SR-ARQ scheme may be found in ISDSN (Integrated Services Digital Satellite Network) [5, 6] where end-to-end error correction is expected to be employed in the future network environment [7, 8].
2. Access Control Strategies — Two priorities queueing systems with access restrictions. The queueing system arises when there are two types of services (e.g. video and data), with one of which has very stringent delay requirement and/or high bandwidth requirement (video or graphics image) while the other (data) vice versa, acquiring access to the network (ISDN). Different “Call level” access control strategies can be applied at the access controller, which is located at the edge of the network. The access control strategies being analyzed include non-preemptive priority scheme and preemptive priority scheme. Both types of service requests in the analysis are queueable and have different service requirements and arrival rates. One of the service types can also acquire simultaneously more than one resources (servers) for service.
3. Modelling of videophone sources and approximation of the effect of high priority packets (e.g. videophone), which usually has correlated arrival process, on

the queueing performance of the low priority packets at the multiplexer. The queueing system arises at the transmission level where the fixed length packets are transmitted at consecutive time slots.

1.3 Highlights of Thesis

Followings are the highlights for each of the topics that has been studied in this thesis :

1. *Analysis of Selective Repeat-Automatic Repeat Request Scheme*

- Derived an exact model for the Select Repeat-Automatic Repeat Request (SR-ARQ) scheme which is applicable for rather small acknowledgment delay time ($N < 7$). (Section 3.3.1)
- Derived an approximate model for the SR-ARQ scheme which is applicable for $N \leq 100$. The result obtained are fairly accurate. (Section 3.3.2)
- Extended the approximate model to analyze the system where the system employs the SR-ARQ scheme and with channel having correlated block errors. (Section 3.5.1)
- Extended the approximate model to analyze the system where the system employs the combination of Forward Error Correction and SR-ARQ scheme. (Section 3.5.2)

2. *Analysis of Access Control Strategies in Wide-band Integrated Network*

- Derived algorithms to solve for the queue length distributions of Two-priority M/M/c queueing systems having heterogeneous arrival and service rates. Both types of services are queueable. Analyses of the following access control strategies are presented :
  - a. Non-preemptive priority policy with high priority services are allowed to have multiple flexible seizures of servers. (Section 4.3)

- b. Non-preemptive priority policy with low priority services are allowed to have multiple flexible seizures of servers. (Section 4.4)
- c. Preemptive priority policy with Type-1 services are allowed to have multiple flexible seizures of servers and preemptive priority to access a designated portion of the servers, and Type-2 services are allowed to have preemptive priority to access the rest of the servers. (Section 4.5)

3. *Videophone Source Modelling and Related Queueing Performance*

- Accomplished a preliminary study on the approximation of the aggregation of videophone sources with a 2-state Markov Modulated Poisson Process. (Section 5.2)
- Derived an approximation for the queueing performance of the low priority packets. The low priority packets' queueing behaviour are under the influence of the high priority packets where the arrival process of the high priority packets is the 2-state Markov Modulated Poisson Process. (Section 5.3.2)

1.4 Outline

The organization of the thesis is described as follows. Chapter 2 introduces the matrix-analytic methodology where the transition matrices and the properties of the basic structured Markov chains or processes are presented. Chapter 3 describes the analysis for the SR-ARQ scheme using an exact and approximate model. Extensions on the approximate model to include correlated block error and hybrid error control scheme are also analyzed. The performance analysis of the various types of access control strategies for two types of prioritized services is given in Chapter 4. Chapter 5 provides the analysis for the videophone source modelling (an example of correlated arrival process for high priority service) and the approximation of the queue length

distribution of the low priority packets under the influence of the high priority packets. Conclusions for the thesis are presented in Chapter 6.

Note 1 : The implementation of the numerical procedures is done via the Mathworks' Matlab package and simulation is implemented using CACI Simscript II.5. general purpose simulation language.

Note 2 : The simulation results shown in this thesis do not include the 95% confidence intervals due to the fact that each simulation run requires approximately 2 days. For the large amount of results and limited time, it is rather difficult to obtain the confidence interval, especially in the case of probability distribution. However, each simulation run consists of large amount of samples and the results obtained may still be considered as acceptable and fairly accurate.

Note 3 : The analytic results presented in the thesis have been checked for validity by comparing with the simulation results. Particularly in the case of the exact analysis, the analytic results shown in the thesis are not accompanied with the simulation results but they have been compared with simulation to ensure the accuracy of the analysis.

Chapter 2

Matrix-Analytic Methodology

2.1 Introduction

Many widely differing stochastic models, such as the ones that are presented in the thesis, have embedded Markov chains with the structure of their transition matrices resemble the structure of those M/G/1 queueing systems or GI/M/1 queueing systems — matrix of Hessenberg form. The matrix-analytic methodology is based upon the particular structures of these types of Markov chain transition matrices such that construction of the algorithmic solutions is made possible. With pure probabilistic arguments and the special block partitioned structure, the invariant probability vector of the Markov chain can be partitioned into small probability vectors and they are interrelated with each other through the matrices which denote the “first passage time distributions”. The significance of this matrix-analytic method lies in the fact that once the boundary probability vectors and the first passage time distribution matrices are found, the rest of the invariant probability vector can be computed readily. Detailed discussions on the matrix-analytic method can be found in [1, 2].

The two main classes of infinite Markov chains, namely “Markov chain of M/G/1 Type” and “Markov chain of GI/M/1 Type”, and some of their important properties will briefly be introduced in this chapter. Two particular structured Markov processes, which

arise in the thesis and are the special cases of the "Markov process of GI/M/1 Type", are also presented.

2.2 Infinite Markov Chain of M/G/1 Type

An infinite Markov chain of M/G/1 type with stationary transition probabilities has block partitioned stochastic matrix $\mathbf{P}$ of the form :

$$\mathbf{P} = \begin{bmatrix} B_0 & B_1 & B_2 & B_3 & \cdots \\ C_0 & A_1 & A_2 & A_3 & \cdots \\ 0 & A_0 & A_1 & A_2 & \cdots \\ 0 & 0 & A_0 & A_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \quad (2.1)$$

where the matrices A_k 's ($k \geq 0$) are $m \times m$ substochastic matrices, B_0 is a $n \times n$ matrix, C_0 is a $m \times n$ matrix and B_k 's ($k \geq 1$) are $n \times m$ matrices. The state space of this Markov chain is the set S :

$$S = \{(0, j), 1 \leq j \leq n, \text{ and } (i, j), i \geq 1, 1 \leq j \leq m\} \quad (2.2)$$

with the matrix $A = \sum_{k=0}^{\infty} A_k$ is a stochastic irreducible matrix and satisfies the equations :

$$\begin{aligned} C_0 \underline{e} + \sum_{k=1}^{\infty} A_k \underline{e} &= \underline{e}, \\ B_0 \underline{e} + \sum_{k=1}^{\infty} B_k \underline{e} &= \underline{e} \end{aligned} \quad (2.3)$$

where $\underline{e} = (1, 1, 1, \dots, 1)^T$.

The steady state probability vector $\underline{x} = [\underline{x}_0, \underline{x}_1, \underline{x}_2, \dots]$ of the matrix $\mathbf{P}$ satisfies

$$\underline{x} \mathbf{P} = \underline{x}, \quad \underline{x} \underline{e} = 1 \quad (2.4)$$

where $\underline{x}_0$ is an n -row vector and $\underline{x}_k$'s, $k \geq 1$ are m -row vectors.

The key steps in solving this type of Markov chain is to determine the following stochastic matrices:

- matrix G : the first passage time distributions from states in the level $\underline{i+1}$ to the level $\underline{i}$ where level $\underline{i}$ refers to the set of states $\underline{i} = \{(i,j), 1 \leq j \leq m\}, i \geq 1$.
- matrices L, H, K : the first passage time distributions from level $\underline{0}$ back to level $\underline{0}$, from level $\underline{1}$ to level $\underline{0}$ and from level $\underline{1}$ to level $\underline{1}$ respectively.

$\underline{x}_0$ and $\underline{x}_1$ can first be obtained from the above matrices (G, L, H , and K), then $\underline{x}_i$'s, $i \geq 2$ can be computed easily with the probability vectors $\underline{x}_0$ and $\underline{x}_1$, the matrix G and the given block matrices in P . Detailed procedures are given in Appendix A.

2.3 Infinite Markov Chain of GI/M/1 Type

Another type of Markov chain of which the transition matrix has transposed Hessenberg form is the Markov chain of GI/M/1 type. The transition probability matrix $\tilde{P}$ of this type of Markov chain is of the form

$$\tilde{P} = \begin{bmatrix} B_0 & A_0 & 0 & 0 & \cdots \\ B_1 & A_1 & A_0 & 0 & \cdots \\ B_2 & A_2 & A_1 & A_0 & \cdots \\ B_3 & A_3 & A_2 & A_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \quad (2.5)$$

in which all elements are $m \times m$ non-negative matrices. For $\tilde{P}$ is stochastic, then

$$B_k \underline{e} + \sum_{v=0}^k A_v \underline{e} = \underline{e}, \quad \text{for } k \geq 0. \quad (2.6)$$

The state space of this Markov chain, S , is

$$S = \{ (i, j), i \geq 0, 1 \leq j \leq m \} \quad (2.7)$$

and the invariant probability vector $\underline{x} = [x_0, x_1, x_2, \dots]$ satisfies

$$\underline{x} \tilde{P} = \underline{x}, \quad \underline{x} e = 1 \quad (2.8)$$

and most importantly,

$$x_{i+1} = \underline{x}_i R, \quad \text{for } i \geq 0 \quad (2.9)$$

where $R = \sum_{k=0}^{\infty} R^k A_k$ is the expected time spent in level $i+1$ before the first return to level i , $i \geq 0$. Here $\underline{x}$ is called the "matrix-geometric invariant probability vector" and R is called the "rate matrix" [1]. For this type of Markov chain, the probability vectors x_i 's, $i \geq 1$ can be computed easily once x_0 and R are found.

The following subsections describe two special cases of the Markov Process of GI/M/1 Type which appear in the stochastic models being studied in this thesis.

2.3.1 Quasi-Birth-and-Death (QBD) Process

A quasi-birth-and-death process is a Markov process on the state space S (as given in Eqn.(2.7)) with infinitesimal generator $\tilde{Q}$ given by

$$\tilde{Q} = \begin{bmatrix} B_0 & D_0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & \dots \\ C_1 & B_1 & D_1 & 0 & \dots & 0 & 0 & 0 & 0 & \dots \\ 0 & C_2 & B_2 & D_2 & \dots & 0 & 0 & 0 & 0 & \dots \\ \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & C_{r-2} & B_{r-2} & D_{r-2} & 0 & 0 & 0 & 0 & \dots \\ 0 & \dots & 0 & C_{r-1} & B_{r-1} & D_{r-1} & 0 & 0 & 0 & \dots \\ 0 & \dots & 0 & 0 & A_2 & A_1 & A_0 & 0 & 0 & \dots \\ 0 & \dots & 0 & 0 & 0 & A_2 & A_1 & A_0 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix} \quad (2.10)$$

where all blocks are $m \times m$ non-negative matrices. The generator $\tilde{Q}$ is irreducible, and for the stable queue, the stationary probability vector

$$\underline{x} = [\underline{x}_0, \underline{x}_1, \dots, \underline{x}_{r-1}, \underline{x}_r, \dots], \quad (2.11)$$

partitioned into m -vectors and satisfies

$$\underline{x} \tilde{Q} = \underline{0}, \quad \underline{x} \underline{e} = 1 \quad (2.12)$$

is again given by

$$\underline{x}_i = \underline{x}_{r-1} R^{i-r+1}, \quad \text{for } i \geq r-1 \quad (2.13)$$

where R (the rate matrix) is the minimal non-negative solution to the equation

$$R^2 A_2 + R A_1 + A_0 = \underline{0} \quad (2.14)$$

and $\underline{x}_0, \underline{x}_1, \dots, \underline{x}_{r-1}$ can be computed from the set of simultaneous matrix equations

$$\begin{aligned} \underline{x}_0 B_0 + \underline{x}_1 C_1 &= \underline{0}, \\ \underline{x}_i D_i + \underline{x}_{i+1} B_{i+1} + \underline{x}_{i+2} C_{i+2} &= \underline{0}, \quad \text{for } 0 \leq i \leq r-3 \\ \underline{x}_{r-2} D_{r-2} + \underline{x}_{r-1} (B_{r-1} + R A_2) &= \underline{0}, \\ \sum_{v=0}^{r-2} \underline{x}_v \underline{e} + \underline{x}_{r-1} (I - R)^{-1} \underline{e} &= 1. \end{aligned} \quad (2.15)$$

Numerical instability is avoided in computing the unique boundary probability vectors $\underline{x}_0, \underline{x}_1, \dots, \underline{x}_{r-1}$ by converting the Markov process to the equivalent embedded Markov chain. Detailed procedures for solving an example of queueing system which involves the QBD process are given in Appendix F.

2.3.2 Markov Process of Quasi-GI/M/1 Type with Complex Boundary behaviour (QGM1-CB process)

The QGM1-CB process is a Markov process on the state space S (given as in Eqn.(2.7)) with infinitesimal generator $\tilde{Q}$ given by

$$\tilde{Q} = \begin{bmatrix} B_{0,0} & B_{0,1} & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots \\ B_{1,0} & B_{1,1} & A_0 & 0 & \cdots & 0 & 0 & 0 & 0 & \cdots \\ B_{2,0} & B_{2,1} & A_1 & A_0 & \cdots & 0 & 0 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ B_{b-1,0} & B_{b-1,1} & A_{b-1} & A_{b-2} & \cdots & A_0 & 0 & 0 & 0 & \cdots \\ 0 & A_{b+1} & A_b & A_{b-1} & \cdots & A_1 & A_0 & 0 & 0 & \cdots \\ 0 & 0 & A_{b+1} & A_b & \cdots & A_2 & A_1 & A_0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix} \quad (2.16)$$

where all blocks are of order m . The generator $\tilde{Q}$ is irreducible. For the stable queue, the stationary probability vector $\underline{x} = [\underline{x}_0, \underline{x}_1, \underline{x}_2, \dots]$, partitioned into m -vectors and satisfies Eqn.(2.12), is given by

$$\underline{x}_i = \underline{x}_1 R^{i-1}, \quad \text{for } i \geq 1 \quad (2.17)$$

where R is the minimal non-negative solution to the equation

$$\sum_{i=0}^{b+1} R^i A_i = 0. \quad (2.18)$$

Boundary probability vectors $\underline{x}_0$ and $\underline{x}_1$ will satisfy

$$(\underline{x}_0 \ \underline{x}_1) \begin{bmatrix} B_{0,0} & B_{0,1} \\ \sum_{i=0}^{b-1} R^i B_{i,0} & \sum_{i=0}^{b-1} R^i B_{i,1} + R^b A_{b+1} \end{bmatrix} = (\underline{0} \ \underline{0}). \quad (2.19)$$

and they can be computed iteratively. Computation of the vector $(\underline{x}_0 \ \underline{x}_1)$ is also done by first converting the system described by Eqn.(2.19) to embedded Markov chain to avoid numerical instability. Detailed procedures for solving an example of queueing system involving the QGM1-CB process are given in Appendix E.

Chapter 3

Queueing Analysis of the Selective Repeat ARQ Scheme

3.1 Introduction

In a data communications system, an error control method must be provided to ensure reliable delivery of data. Automatic repeat request (ARQ) schemes have long been utilized to control errors on two-way digital transmission links. For the system having a feedback channel, i.e. a channel for the transmission of positive / negative acknowledgment (ACK / NACK), various ARQ schemes have been studied [9]. The selective-repeat ARQ (SR-ARQ) scheme is known to have the highest throughput among the pure ARQ schemes. The queueing analysis of the SR-ARQ scheme presents a computationally difficult problem when the ACK time, i.e. the time from the transmission of a block of data to the reception of an ACK or NACK for the block, is long compared to block transmission time [3, 4]. The difficulty is due to the fact that the complexity of the algorithm increases exponentially with ACK time.

Studies have indicated that in a high-speed communications network environment, *edge-to-edge* error control scheme (i.e. error recovery is performed at the edge of the network) is more suitable than the conventional *link-by-link* error control scheme [7, 8]. It is conjectured that the ACK time in the edge-to-edge error control scheme will

considerably be longer than the transmission time and hence previous exact analysis on the SR-ARQ scheme may not be applicable. An approximate model which totally ignores the dependence of NACK arrivals on the past history of block transmissions has been proposed for a system with a long ACK time [4]. Although the model can be used to predict queueing behaviour under a wide range of parameter values, the result of the approximation is not in general satisfactory. The model fails to provide a good approximation when it is compared with exact analysis or with simulation results under certain conditions, even if the ACK time is rather long. Analysing system which has a slightly complicated error control scheme or a system with correlated block errors is even more difficult with the classical analytic methods. The difficulty stems from the increase in dimensionality, which is caused by the extra description required for the model.

In this chapter, the queueing behaviour of the SR-ARQ scheme is analyzed using the matrix-analytic approach [1, 2]. Even with the application of matrix-analytic method, exact analysis is practical only when the ACK time is short (in spite of the fact that this ACK time limit has been extended when it is compared with the previous analysis). For a system with a long ACK time, an approximate model which does take the dependence of NACK arrivals on the past history of block transmissions into considerations is proposed and analyzed. Following is the outline for this chapter : in section 3.2, the basic SR-ARQ protocol, the queueing model and the assumptions for the analysis are presented. The queue length distribution using an exact model and an approximate model is analyzed in section 3.3. Numerical examples for the exact and the approximate model are presented in section 3.4. In section 3.5, two possible extensions of the approximation to the analysis of the system with *correlated* block errors and the system using a *hybrid* of *forward error correction* and *SR-ARQ* (FEC/SR-ARQ) scheme are presented. Conclusions of the queueing analysis for the SR-ARQ schemes are given in section 3.6.

3.2 The Model

3.2.1 Basic system and the SR-ARQ protocol

The system under consideration consists of a *sending node* connected to a *receiving node* through a logical link with the application of edge-to-edge error control scheme (i.e. system with long ACK time). The nodes can be viewed essentially as (slotted) concentrator network nodes which are located at the edge of the network. The *sending node* sends data blocks to the *receiving node*. The *receiving node* is required to make a (positive/negative) acknowledgment (ACK/NACK) to the *sending node* (see figure 3.1). Until an acknowledgment is received, a “copy” of the data block must be retained at the *sending node*. If an ACK is received, the space assigned to the “copy” is released. Meanwhile, the *sending node* sends the block next in the queue as long as there are blocks waiting to be sent. If either a NACK or no acknowledgment is received in a suitable interval of time (time-out period), the *sending node* immediately retransmits the block, which is received in error or has no acknowledgment, with absolute priority. The *receiving node* is responsible for rearranging the blocks if necessary.

Figure 3.1 Basic system for the SR-ARQ scheme

3.2.2 Assumptions for the model

The following assumptions for the model are made (see figure 3.2) :

1. Messages (data) are divided into fixed length blocks.

Figure 3.2 Queueing model for SR-ARQ scheme

2. Channel time is divided into contiguous slots of a block transmission time (i.e. a slotted channel). Block transmissions are synchronized to the slots.
3. The ACK time (and the time-out period) is N slots, i.e. the time between the beginning of a block transmission and the reception of an ACK or NACK for the block at the sending node is N slots.
4. Blocks of data arrive and join the *transmit queue* at the sending node. Once transmitted, a copy of the block is stored at the *transit queue* until an ACK for the block is received. When a NACK for a block is received, the copy of the block is placed at the head of the transmit queue for retransmission. The size of the transmit queue is *infinite*.
5. The blocks arrive just prior to a slot. Denote Y_n as the number of blocks arriving at the end of the n^{th} slot. Y_n 's are independent and identically distributed with

$$y_k = \Pr[Y_n = k], \quad (k \geq 0, \text{ and } n \geq 1) \quad (3.1)$$

The mean number of blocks arriving in one slot is denoted by λ , i.e., $\lambda = E[Y_n]$. ($n \geq 1$).

6. The event that a block is received in error in a slot is independent of those in the other slots. The followings are defined :

$$\alpha = \Pr [\text{a block is received in error}], \quad \beta = 1 - \alpha. \quad (3.2)$$

* This assumption will be relaxed later in the extended model (see section 3.5).

3.3 Analysis

The system at the beginning of the n^{th} slot can be described by the state vector

$$\underline{S}_n = (Q_n, L_{n,0}, L_{n,1}, L_{n,2}, \dots, L_{n,N-1}), \quad (3.3)$$

where

Q_n = the number of blocks in *transmit* queue.

$$L_{n,i} = \begin{cases} 1, & \text{if a block was transmitted } (i+1) \text{ slots before,} \\ 0, & \text{otherwise.} \end{cases} \quad (3.4)$$

The sequence $\{\underline{S}_n, n \geq 1\}$ is a first-order Markov chain whose state space is $\bigcup_{Q_n=0}^{\infty} \bigcup_{L_{n,0}=0}^1 \dots \bigcup_{L_{n,N-1}=0}^1 (Q_n, L_{n,0}, L_{n,1}, \dots, L_{n,N-1})$. The size of the state space grows in proportion to 2^N .

3.3.1 Matrix-analytic approach (an exact analysis)

The exact position of the slots in the transit queue being occupied by data blocks is recorded with Eqn.(3.3) and Eqn.(3.4), which is equivalent to a binary representation. Knowing the equivalent decimal representation of the binary sequence will immediately indicate the status of transit queue slot occupancy. By defining

$$L_n = \sum_{i=0}^{N-1} L_{n,i} 2^i, \quad (3.5)$$

then $\{\underline{S}_n = (Q_n, L_n), n \geq 1\}$ is a two-dimensional Markov chain. Note that when $Q_n \geq 1$, $\Pr[L_n = 2^l] = 0$ (l is a non-negative integer) since $L_{n,0} = 1$ (i.e. the first slot in the transit queue will always be occupied whenever there are data blocks waiting in transmit queue). Let $m = 2^{N-1}$ and define the steady state transmit queue length probabilities as

$$x_{k,l} = \begin{cases} \Pr[Q_n = k, L_n = l], & k = 0, 0 \leq l \leq 2m - 1, \\ \Pr[Q_n = k, L_n = 2l + 1], & k \geq 1, 0 \leq l \leq m - 1. \end{cases} \quad (3.6)$$

Define the steady state probability vectors as

$$\begin{aligned}\underline{x}_k &= (x_{k,0}, x_{k,1}, \dots, x_{k,m-1}), \quad k \geq 1, \\ \underline{x}_0 &= (x_{0,0}, x_{0,1}, \dots, x_{0,2m-1}),\end{aligned}\tag{3.7}$$

and

$$\underline{x} = (\underline{x}_0, \underline{x}_1, \underline{x}_2, \dots),\tag{3.8}$$

then,

$$\underline{x} \mathbf{P} = \underline{x}, \quad \underline{x} \underline{e} = 1,\tag{3.9}$$

where

$$\underline{e} = (1, 1, 1, 1, \dots)^T,\tag{3.10}$$

and the transition probability matrix $\mathbf{P}$ is given by

$$\mathbf{P} = \begin{bmatrix} B_0 & B_1 & B_2 & B_3 & \dots \\ C_0 & A_1 & A_2 & A_3 & \dots \\ 0 & A_0 & A_1 & A_2 & \dots \\ 0 & 0 & A_0 & A_1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.\tag{3.11}$$

B_0 is a $2m \times 2m$ matrix, C_0 is a $m \times 2m$ matrix, B_k 's ($k \geq 1$) are $2m \times m$ matrices and A_k 's ($k \geq 0$) are $m \times m$ matrices. The elements of the submatrices are given by

$$B_0(i, j) = \begin{cases} y_0, & 0 \leq i \leq m-1, j = 2i \\ y_1, & 0 \leq i \leq m-1, j = 2i+1 \\ \beta y_0, & m \leq i \leq 2m-1, j = 2(i-m) \\ \beta y_1 + \alpha y_0, & m \leq i \leq 2m-1, j = 2(i-m)+1 \\ 0, & \text{otherwise} \end{cases}\tag{3.12}$$

for $0 \leq i, j \leq 2m-1$,

$$B_k(i, j) = \begin{cases} y_{k+1}, & 0 \leq i \leq m-1, j = i \\ \beta y_{k+1} + \alpha y_k, & m \leq i \leq 2m-1, j = i-m \\ 0, & \text{otherwise} \end{cases}$$

for $k \geq 1, 0 \leq i \leq 2m-1$ and $0 \leq j \leq m-1$,

$$\begin{aligned}
C_0(i, j) &= \begin{cases} y_0, & 0 \leq i \leq m/2 - 1, j = 4(i + 1) - 1 \\ \beta y_0, & m/2 \leq i \leq m - 1, j = 4(i - m/2 + 1) - 1 \\ 0, & \text{otherwise} \end{cases} \\
&\text{for } 0 \leq i \leq m - 1 \text{ and } 0 \leq j \leq 2m - 1, \\
A_0(i, j) &= \begin{cases} y_0, & 0 \leq i \leq m/2 - 1, j = 2i + 1 \\ \beta y_0, & m/2 \leq i \leq m - 1, j = 2(i - m/2) + 1 \\ 0, & \text{otherwise} \end{cases} \\
&\text{for } 0 \leq i, j \leq m - 1, \\
A_k(i, j) &= \begin{cases} y_k, & 0 \leq i \leq m/2 - 1, j = 2i + 1 \\ \beta y_k + \alpha y_{k-1}, & m/2 \leq i \leq m - 1, j = 2(i - m/2) + 1 \\ 0, & \text{otherwise} \end{cases} \\
&\text{for } k \geq 1 \text{ and } 0 \leq i, j \leq m - 1.
\end{aligned} \tag{3.13}$$

The transition probability matrix $\mathbf{P}$ shows that the Markov chain at hand is of “M/G/1 type” [1]. Following Lucantoni’s approach, one can readily obtain the probability vectors $\underline{x}_0$ and $\underline{x}_1$ [2]. The remaining probability vectors, $\underline{x}_k$ ’s ($k \geq 2$), are then obtained using Ramaswami’s method [10]. In Appendix A, the algorithm used in the numerical examples is presented.

3.3.2 Approximate model and its analysis

Since the size of the submatrices grows approximately with 2^{N-1} , the exact analysis is limited to a rather small value of N (i.e. the ACK time). The computational experience shows that the exact analysis becomes impractical for $N > 7$. In the approximate model, a less detailed system states is used by re-defining

$$J_n = \sum_{i=0}^{N-1} L_{n,i}, \tag{3.14}$$

where L_n now is the total number of blocks in the *transit* queue. The following *approximate* assumption is made.

Assumption : Given that $L_n = l$, the probability of the event $[L_{n,N-1} = 1]$ is

$$\Pr[L_{n,N-1} = 1 \mid L_n = l, Q_n = k] = \begin{cases} \frac{\binom{N-1}{l-1}}{\binom{N}{l}} = \frac{l}{N}, & \text{if } k = 0 \\ \frac{\binom{N-2}{l-2}}{\binom{N-1}{l-1}} = \frac{l-1}{N-1}, & \text{if } k \geq 1 \end{cases} \quad (3.15)$$

and the event is independent of the events $[L_{i,N-1} = 1]$, ($i < n$).

This assumption essentially ignores the exact position of the data blocks but retains only the information about the total number of data blocks in the transit queue. The last slot of the transit queue being occupied by a data block is then random from one time slot to the next time slot. The approximation by Anagnostou *et. al.* [4] essentially uses a model without the feedback channel, where their assumption is that the retransmission requests (NACK) arrive at the sending node independently of the past history of the system. Under heavy load, their approximation tends to fail. Since when the system is under a heavy load, the probability that the last slot being occupied by a data block is high, while under a light load, this probability is low and does not influence the queueing behaviour of the transmit queue significantly, hence the assumption made in this proposed approximation can be justified.

Due to the approximate assumption, the sequence $\{\underline{S}_n = (Q_n, L_n), n \geq 1\}$ forms a 2-dimensional first-order Markov chain. Here, the probability $x_{k,l}$ is defined as

$$x_{k,l} = \begin{cases} \Pr[Q_n = k, L_n = l], & k = 0, 0 \leq l \leq N, \\ \Pr[Q_n = k, L_n = l + 1], & k \geq 1, 0 \leq l \leq N - 1. \end{cases} \quad (3.16)$$

and

$$\begin{aligned} \underline{x}_k &= (x_{k,0}, x_{k,1}, \dots, x_{k,N-1}), \quad k \geq 1, \\ \underline{x}_0 &= (x_{0,0}, x_{0,1}, \dots, x_{0,N}), \end{aligned} \quad (3.17)$$

and

$$\underline{x} = (\underline{x}_0, \underline{x}_1, \underline{x}_2, \dots). \quad (3.18)$$

The transition probability matrix $\mathbf{P}$ is again given by Eqn.(3.11) where B_0 is a $(N+1) \times (N+1)$ matrix, C_0 is a $N \times (N+1)$ matrix, B_k 's ($k \geq 1$) are $(N+1) \times N$ matrices and A_k 's ($k \geq 0$) are $N \times N$ matrices. The elements of the submatrices are given by

$$\begin{aligned}
 B_0(i, j) &= \begin{cases} y_1(1 - \frac{1}{N}), & 0 \leq i \leq N, j = i + 1 \\ y_0(1 - \frac{1}{N} + \alpha \frac{1}{N}) + \beta y_1 \frac{1}{N}, & 0 \leq i \leq N, j = i \\ \beta y_0 \frac{1}{N}, & 0 \leq i \leq N, j = i - 1 \\ 0, & \text{otherwise} \end{cases} \\
 &\text{for } 0 \leq i, j \leq N, \\
 B_k(i, j) &= \begin{cases} y_{k+1}(1 - \frac{1}{N}), & 0 \leq i \leq N, j = i \\ \beta y_{k+1} \frac{1}{N} + \alpha y_k \frac{1}{N}, & 0 \leq i \leq N, j = i - 1 \\ 0, & \text{otherwise} \end{cases} \\
 &\text{for } k \geq 1, 0 \leq i \leq N \text{ and } 0 \leq j \leq N - 1, \\
 A_0(i, j) &= \begin{cases} \beta y_0(\frac{1}{N-1}), & 0 \leq i \leq N - 1, j = i \\ y_0(1 - \frac{1}{N-1}), & 0 \leq i \leq N - 1, j = i + 1 \\ 0, & \text{otherwise} \end{cases} \\
 &\text{for } 0 \leq i, j \leq N - 1, \\
 A_k(i, j) &= \begin{cases} (\alpha y_{k+1} + \beta y_k) \frac{1}{N-1}, & 0 \leq i \leq N - 1, j = i \\ y_k(1 - \frac{1}{N-1}), & 0 \leq i \leq N - 1, j = i + 1 \\ 0, & \text{otherwise} \end{cases} \\
 &\text{for } k \geq 1 \text{ and } 0 \leq i, j \leq N - 1.
 \end{aligned} \tag{3.19}$$

and $C_0 = [\underline{0} \ A_0]$, where $\underline{0} = (0, 0, \dots, 0)^T$.

Note : The (i, j) element of the above matrices is the probability that the number of data blocks in the transit queue observed just after the transmission of the slot changes from u data blocks to v data blocks in one slot where

- $u = i, v = j$ for matrix B_0 ,
- $u = i, v = j + 1$ for matrix B_k 's, ($k \geq 1$),

- $u = i + 1, v = j$ for matrix C_0 ,
- $u = i + 1, v = j + 1$ for matrix A_k 's, ($k \geq 0$).

The steady state distribution $\underline{x}$ can be obtained by the same method used in the exact analysis (see Appendix A).

3.4 Numerical Examples for SR-ARQ scheme

Probability distribution of Q_n is plotted in Figures 3.3 to 3.6 under various block arrival rates λ , block error rates α , and ACK times N . It should be noted that the total number of blocks buffered at the sending node ($Q_n + L_n$) can also be easily obtained once $\underline{x}$ is computed. The results obtained by the approximate method (Approximation 1 in the figures) are compared with the exact analysis for $N \leq 7$ and with the simulations for $N > 7$. The approximation by Anagnostou *et. al.* [4] (Approximation 2 in the figures) is also included for comparison. It should be noticed that the proposed approximation and the exact analysis are almost identical as shown in Figures 3.3 and 3.4 — the solid line (exact) and the dashed line (Approximation 1) are so close that they appear as one line. These comparisons show that the proposed approximation provides accurate estimation of queue length distribution under a wide range of conditions. The approximation by Anagnostou *et. al.* also provides accurate results for extremely large value of N , moderate arrival rates or rather small block error rates. However, if the channel utilization is high and ACK time is moderate (e.g. see Figure 3.6 for $\lambda = 0.65$, $\alpha = 0.3$ and $N = 100$), their method significantly under-estimates the queue size. Throughout the numerical examples, Y_n is assumed to be Poisson distributed. Each simulation curve reported here is the result on at least 1,000,000 slots of simulation.

Computational Experience — On a Sun 3/60 20Mhz machine with a share of 30% cpu time, the average computation time for one set of parameter value is about 45

minutes to 1 hour. When the total system utilization is high (approximately 0.7 or higher) and N is large (approximately 50), the computation time is sometimes 3 to 6 hours. If the numerical procedures are implemented using C or Fortran instead of the pro-Matlab package, the computation time is expected to be 5 to 7 times less. In order to obtain a reliable simulation result under high utilization and large N , simulation time for one set of parameter is approximately 18 to 22 hours using the Simscript II.5 simulation language.

3.5 Extension of the Approximate Model

3.5.1 Channel with correlated block errors

In a system with a fading channel where the duration of a fade is long compared with the block transmission time, the occurrences of block errors tend to be correlated. Similar can be said in a network under a congestion during which blocks are lost frequently if the duration of a congestion period is long compared with the block transmission time. The approximate model can be extended for the analysis of the SR-ARQ scheme in the system with correlated block errors, for example, the satellite networks and packet radio.

The channel used in the analysis is assumed to be in one of the two possible states, *good* (G) or *bad* (B). It is further assumed that the channel states form a two-state discrete time Markov chain (see Figure 3.7). The transition probabilities of the channel state is given by

$$\begin{aligned} a &= \Pr[D_n = B \mid D_{n-1} = G] \\ b &= \Pr[D_n = G \mid D_{n-1} = B] \end{aligned} \tag{3.20}$$

where $D_n \in \{G, B\}$ is the channel state during the n^{th} slot. Block error rate is α_G when $D_n = G$, and it is α_B when $D_n = B$. The average block error rate is given by $\alpha = (b\alpha_G + a\alpha_B) / (a + b)$.

Figure 3.3 Transmit queue length distribution ($N = 5$, $\alpha = 0.01$)

Figure 3.4 Transmit queue length distribution ($N = 5$, $\alpha = 0.1$)

Figure 3.5 Transmit queue length distribution ($N = 10$, $\alpha = 0.3$)

Figure 3.6 Transmit queue length distribution ($N = 100$, $\alpha = 0.3$)

Figure 3.7 Channel model with correlated block error rates

The system can be described by $\underline{S}_n = (Q_n, L_n, D_n)$, $(n \geq 1)$. Due to the approximate assumption (Eqn.(3.15)), the sequence of $\underline{S}_n$'s forms a first-order Markov chain. Once the transition probability matrix of the chain, $\mathbf{P}$, is found, the steady state transmit queue length distribution can be obtained using the algorithm in Appendix A. The method can be generalized for the system with more than two channel states. In Appendix B, the elements of $\mathbf{P}$ for the system with an arbitrary number of channel states are presented.

3.5.2 System with FEC/SR-ARQ schemes

The use of forward error correction (FEC) in conjunction with an ARQ scheme has been proposed for a noisy channel [11, 12]. The hybrid scheme (FEC/SR-ARQ scheme, or *modified memory ARQ* in [11]) reduces the probability of repeated request for retransmissions. An erroneous block is not discarded at the receiver. The subsequent transmission is not just a repetition of the previous transmission, but contains carefully structured incremental redundancy in such a way that alternate transmissions together form a *rate 1/2 code*. Using two consecutive transmissions for the same message, the receiving node is able to restore the original message with a higher probability of success. The hybrid scheme results in an increased maximum throughput and a decreased receiver buffer size requirement [11].

It is possible to analyze the queue length distribution at the sending node for a hybrid FEC/SR-ARQ scheme in which the probability of the first retransmission request of a

block, α_F , and the probability of the subsequent retransmission request of a block (after one or more previous transmission attempts), α_R ($\leq \alpha_F$), are different. The system can be *approximately* described by the state $\underline{S}_n = (Q_n, L_n, \tilde{L}_{n,R})$ where

$\tilde{L}_{n,R}$ = total number of blocks in the *transit* queue which are retransmitted. (3.21)

The sequence $\{\underline{S}_n, n \geq 1\}$ forms a first-order Markov chain under the following *approximate* assumption:

$$\Pr[L_{n,N-1} = R \mid \tilde{L}_{n,R} = l, L_n = i, Q_n = k] = \frac{l}{N}, \quad (3.22)$$

$$\Pr[L_{n,N-1} = F \mid \tilde{L}_{n,R} = l, L_n = i, Q_n = k] = \frac{i-l}{N}, \quad (3.23)$$

where $L_{n,N-1} = F$ (or R) denotes that a block was transmitted for the first time (or retransmitted) during the $(n-(N-1))^{\text{th}}$ slot. Note that in equations (3.22) and (3.23), the fact that $L_{n,0} = 1$ when $Q_n \geq 1$ is ignored in order to simplify the analysis. The method of obtaining the steady state queue length distribution is the same as the one used in the previous approximation. The elements of the transition probability matrix of the chain, $\mathbf{P}$, of this model is given in Appendix C.

The expected number of transmissions of a block until a success occurs is given by

$$(1 - \alpha_F) + \alpha_F(1 - \alpha_R) \sum_{j=0}^{\infty} (j+2) \alpha_R^j = 1 + \frac{\alpha_F}{1 - \alpha_R} \quad (3.24)$$

and the average block error rate is therefore given by $\alpha = \alpha_F / (1 + \alpha_F - \alpha_R)$.

3.5.3 Numerical Examples

Figure 3.8 shows a sample numerical result for the following parameters :

- i. $\lambda = 0.65$
- ii. $N = 10$
- iii. $\alpha_G = 0.15, \alpha_B = 0.45, a = b = 0.1$ (or $a = b = 0.05$)
- iv. $\alpha_F = 0.3, \alpha_R = 0.15$ (or $\alpha_F = 0.35294, \alpha_R = 0.17647$).

It should be noticed that the approximate analysis gives reasonably accurate transmit queue length distribution compared with simulation results.

Channel with a stronger positive correlation (smaller value of a and b) results in a considerable increase in queue length. As expected, the queue length of the hybrid scheme is substantially reduced compared with the SR-ARQ scheme without FEC. In order to compare the SR-ARQ scheme with and without FEC under the same channel utilization $\rho = \lambda \times (\text{expected number of transmission of a block})$, α_F is increased from 0.3 to 0.35294 and α_R from 0.15 to 0.17647 respectively. The resulting transmit queue length distribution is also plotted in Figure 3.8. It is observed that even under the same utilization, the hybrid scheme has a somewhat smaller queue occupancy.

3.6 Conclusions

Previous work on the queueing analysis of SR-ARQ scheme is limited for small ($N \leq 4$) or extremely large ACK times. It appears no way of obtaining an *exact* solution without the consideration of the number of states (corresponding to the history) which grows exponentially with N . This limits severely the applicability of the solution, for example in the evaluation of a satellite system employing SR-ARQ scheme or a high-speed communications network employing edge-to-edge SR-ARQ scheme in which N will be considerably large ($N > 7$). In this chapter, the (transmit) queue length distribution of SR-ARQ using matrix-analytic approach is analyzed. With the matrix-analytic approach, the dimensionality constraint (2^N) is relaxed from the previous analysis of 2^4 to 2^7 , but $N = 7$ may still not be high enough for the analysis to be useful. An approximate model based on partial history of the system is hence proposed and analyzed. Using the approximate model, an accurate queue length distribution of SR-ARQ scheme can be obtained under the condition in which the previously known methods fail to provide reasonable results or are difficult to apply due to numerical

Figure 3.8 Transmit queue length distribution for the system with correlated block errors and the system using hybrid FEC/SR-ARQ scheme ($N=10$)

problems. The ACK time N can now be extended to approximately 100 time slots using the proposed approximate model. Possible extensions to the approximate model are also presented — the system with correlated block errors or hybrid of FEC/SR-ARQ scheme can be analyzed without incurring undue increase in numerical complexity.

Note: In the future Broadband-ISDN, the cell loss rate within the ATM (asynchronous transfer mode) network is anticipated to be very low (approximately 10^{-10}). The approximation by Anagnostou *et. al.* seems to be appropriate for such situation. However, error control is expected to be performed at the edge of the network and the user may not treat individual ATM cell as a unit but rather a collection of ATM cells as single unit, the overall error rate experienced by the end-user may be high. If an end user considers a collection of ATM cells as approximately 100 ATM cells and each ATM cell is approximately 50 bytes, then the error rate of a collection of ATM cells may become 5^{-7} . In addition, other errors such as propagation error and fluctuation in channel state may contribute to the total error rate. The overall error rate expected may become 10^{-5} . As compared to the parameter values chosen in the numerical examples, error rates of 0.3, 0.1 or 0.01 are very unlikely to occur under normal operation of the ISDN. However, for a packet radio system operating in the realm of ISDN, the error rate may sometimes be high. For example, when the system is operating in a thunder storm, the error rate will be high and this analysis may become useful in that particular situation. One of the purposes of using the error rate of 0.1 or 0.3 in the numerical examples is to demonstrate the accuracy of the approximation even under high error rate.

In addition, FEC/SR-ARQ scheme may seem to be unnecessary within the network due to the small error rate anticipated, but from the end-user's point of view, FEC/SR-ARQ scheme may not totally be obsolete. The analysis provided in this chapter may hence provide certain insight as to how the various types of error control schemes can be analyzed under situation where error rate is not small.

Chapter 4

Queueing Analysis of Traffic Access Control Strategies in Wide-Band Integrated Networks

4.1 Introduction

The field of modern telecommunications is evolving towards the realm of wide-band Integrated Services Digital Network (ISDN) where access rates will be capable of supporting services beyond 100 Mbps. The ISDN provides a common transmission facility to serve a wide variety of user needs and the traffic demands from these users vary in their arrival rates, service rates and bit rates. Moreover, traffics will differ considerably in their signal quality (such as in video/audio), access throughput, and admissible delay, etc. In order to allocate the available capacity to the various types of traffics in an efficient, flexible and fair manner, a channel access control strategy at "call admission level" is required. Access is defined here as the link from the service-request (SR) premises to the first ISDN entrance node. The service-request premises may consist of a set of traffic generating sources such as telephone sets, facsimile sets and video sets, etc. The ISDN entrance node is the servicing centre capable of providing various types of switching and multiplexing operations with sufficient amount of control intelligence. A simplified view of the ISDN access structure is given in Fig.4.1.

Figure 4.1 A simplified view of the structure of ISDN access

4.1.1 Channel Structure

A common high-speed transmission facility (digital pipe/channel — such as optical link) and serving two types of traffics (a narrow-band (NB) traffic and a wide-band (WB) traffic) is considered in the study. The access node of the network, which is available on the service-request (SR) premises, feeds into the ISDN through the high-speed channel. Access control is performed at the access node controller to regulate smoothly and fairly the access of the different types of demands onto the ISDN resources. For convenience, service-request (SR) of WB type and NB type shall hereby be referred to as Type 1–SR and Type 2–SR respectively.

The digital channel's capacity is assumed to consist of c basic bandwidth units (BBUs). If the facility transmits information using time division multiplexing (TDM), then the bandwidth resource can be represented by a frame of c identical slots (servers) where service-requests demand one or more slots (servers) for the duration of their service times (see Fig.4.2).

4.1.2 Integrated Access Control Problem and Related Works

Absence of control in multiplexing streams of heterogeneous traffics may lead to degradations in performance due to the stringent requirements of the various types of

Figure 4.2 Control of integrated access at the ISDN entrance node

traffic demands (service-requests). Classes of restricted access strategies have been proposed and studied in some previously related works. Before introducing the contribution of this study, previous works in this area are briefly reviewed. Note that in the numerical implementations of most of the previous studies, two types of service-requests (Type 1 (WB)-SR and Type 2 (NB)-SR) are considered and the analysis is mainly based on z-transform technique and probabilistic arguments. For convenience and clarity in presenting the previous works, the following notations and assumptions shall be adopted first : Type 1-SR and Type 2-SR have heterogeneous exponential service times and Poisson arrival rates. Type 1-SR requires b , $b \leq c$, servers and Type 2-SR requires only one server out of c servers for service.

Four access control strategies, which are described as two-side moving boundary strategies with preemptive priority (PP) and combination of blockable/queueable policies, are studied in [13] to handle a combination of WB and NB service requests over a common TDM channel. In their analysis, Type 1-SP is either blocked or queued (according to the policies) whenever there are less than b servers available upon its arrival at the system. Transmission capacity is wasted since there is server idle while

there are still Type 1-SRs pending in the queue (if they are queueable according to the policy).

In [14, 15], Type 1-SR is queued, and Type 2-SR is blocked and lost when the respective SR arrives at the system and found that there is not enough server available. In their analysis, Type 1-SR cannot enter service unless at least b servers are available and the c servers are partitioned so that Type 1-SR only have access to a portion of the c servers to avoid unfairness for the Type 2-SR. Transmission capacity is also wasted with this strategy.

A preemptive priority (PP) access control strategy is proposed in [16] by Kraimeche and Schwartz where c servers are partitioned into 2 groups (WB group and NB group). Basically, WB traffic (Type 1-SR) is blocked and lost while NB traffic (Type 2-SR) is queued when no server is available upon arrival. In their model, Type 1-SR still requires b servers for service, but whenever there is server available in the NB group upon a Type 1-SR's arrival at the system, that particular Type 1-SR can be served on the condition that it will have its service time and bit rate requirement (number of servers required) completely transformed to become a Type 2-SR. A Type 2-SR can also use the WB group if there is server available, and it is preempted whenever a Type 1 (WB) SR arrives at the system. Slight modification on this model and its analysis are reported in [17] by Kraimeche and Schwartz where the differences are that Type 1-SR and Type 2-SR are both queueable in [17], and Type 2-SR cannot use WB group. No preemption is necessary since Type 2-SR will be queued if no server in NB group is available without regard to the availability of WB group servers.

In another PP (preemptive-priority) policy of Kraimeche and Schwartz [18], both types of SRs (service-requests) are queueable, i.e. the arriving SRs finding insufficient number of servers available are queued until servers become available. Type 1-SR have preemptive priority over Type 2-SR. Type 2-SR, therefore, use the remaining capacity

left after Type 1–SR are served. In order to prevent the monopolization of the servers by Type 1–SR's, Kim and Un [19] imposes a restriction that Type 1–SR cannot preempt Type 2–SR when the number of servers used by Type 1–SR is greater than or equal to c_1 ($\leq c$). The blocked Type 1–SR are assumed to be lost (non-queueable) in reference [19].

4.1.3 Objectives and Outlines

In the previous works, Type 1–SR will be blocked and lost or queued when it arrives to the system finding insufficient number of servers ($< b$) available for its service. Although compression technique is assumed in [16, 17], where Type 1 service has its bit rate and service time reduced to conform to Type 2 service, to minimize the waste of capacity, yet sometimes this may not be desirable. It would sometimes be appropriate for the Type 1 service to use the available servers ($\leq b$) and have the duration of service time adjusted according to the number of BBU that it can obtain, i.e. if the number of available servers is less than b , then the service will proportionally have longer duration to compensate for the lack of BBUs, and have the priority to resume the service at its original rate (bandwidth) upon the releasing of servers by other SRs. Besides, most of the SRs are either blocked and lost or queued but not both queueable. There may arise situation where the user would like to have the service requests be pended at the access node instead of requesting resources for service in a “trial and error” manner. In addition, non-preemptive priority (NPP) policy is sometimes desirable over preemptive priority (PP) policy to reduce the complexity of the control mechanism of the access node controller since the controller does not have to keep track of the records and status of the services that are being preempted. In order to avoid unfairness caused by the monopolization of servers by a single type of SRs, the c servers may be divided into two groups (NB group and WB group) such that Type 1–SRs have the priority to access the WB group of servers while Type 2–SRs have priority to access the NB group.

In this chapter, different access control strategies (systems employing NPP and PP policies) are studied via the matrix analytic method. Since the analysis for system employing NPP policy with arbitrary number of servers designated as NB or WB group is very difficult to obtain, only special cases are analyzed. The special cases are the boundary cases where a) all servers are designated as WB group, and b) all servers are designated as NB group. The three access control strategies studied allow both Type 1– and Type 2–SR be queued in the respective queues and Type 1–SR will be serviced with its service time adjusted according to the number of servers granted to it. The first strategy assumes that Type 1–SR has absolute non-preemptive priority over Type 2–SR, i.e. Type 1–SR will not preempt any Type 2 service that is occupying the resources (servers) but it has the priority over Type 2–SR to enter service when servers are available. The second strategy is similar and the difference is that Type 2–SR has absolute non-preemptive priority over Type 1–SR. In the third strategy, Type 1–SRs have preemptive priority over Type 2–SR in c_1 ($\leq c$) servers (or WB group) only and that Type 2–SR have preemptive priority over Type 1–SR in the remaining c_2 ($= c - c_1$) servers (or NB group). It is noted that unlike in reference [19], Type 2–SR is allowed to preempt Type 1–SR when there are more than c_1 servers serving Type 1–SR and that the blocked or preempted Type 1–SR are queued instead of being discarded.

Outlines of this chapter are given as follows. Section 4.2 describes the model used and the necessary assumptions, which are common to all three strategies. Section 4.3, 4.4, and 4.5 present the analysis of the three strategies described in the previous paragraph with respective order. Section 4.6 presents numerical examples on the performance of the three types of strategies. Conclusions are given in section 4.7.

4.2 The Model

It is assumed that a common digital transmission capacity is divided into c basic

bandwidth units (BBUs). The WB traffic (Type 1-SR) requires a maximum of b , $b \leq c$ BBUs, and the NB traffic (Type 2-SR) requires one BBU. Type i -SR ($i = 1, 2$) arrives according to a Poisson process with arrival rate λ_i . Service time is exponentially distributed with mean service time of μ_i^{-1} for Type i -SR ($i = 1, 2$), and all BBUs which are assigned to a Type 1-SR during the service is released together.

Both Type 1-SR and Type 2-SR are queueable in the infinite queues with respect to their types. Whenever there are BBUs available for Type 1-SR, which may be less than required (i.e. $< b$), Type 1 service will use the available BBUs ($\leq b$) and the duration of service time will proportionally be adjusted according to the number of BBU that it obtains, i.e. if the number of available BBUs is d ($0 < d \leq b$), then the service rate will be $d\mu_1/b$ to compensate for the lack of BBUs. The Type 1-SR, which is already in service but not yet acquired the maximum number of BBUs that it is allowed, will have the priority over those Type 1-SR pending in the queue to seize the available BBUs which are released by other SRs upon their finishing of services. The system being considered can be modelled as an M/M/c two-priority queueing system with heterogeneous arrival and service rates. The priority in assigning the available BBUs (servers) to the two types of services is governed by the policy being used, and this will be discussed in the subsequent sections. Fig.4.3 illustrates the basic queueing system that is being considered.

4.3 Analysis 1 — Non-Preemptive Priority M/M/c Model : **Type 1-SR has priority over Type 2-SR**

In this model, Type 1-SR has absolute non-preemptive priority over Type 2-SR. When any server is occupied by Type 2-SR, Type 1-SR cannot preempt. When any server is released upon the departure of any SR, Type 1-SR has priority over Type 2-SR in seizing the servers for service. In [20], Miller has presented an analysis for an M/M/c

two-priority queueing system with heterogeneous rates via the matrix analytic method. In his analysis, Type 1-SR has non-preemptive priority over Type 2-SR, but both Type 1- and Type 2-SR only require one server for service. The analysis presented in this section is a generalization of his work where Type 1-SR can seize a maximum of b servers for service and it can enter service even when the number of servers available are less than b , $b \leq c$. One of the difficulties in solving this type of non-preemptive queueing system is due to the infinite dimension caused by infinite queue size of both Type 1 and Type 2 queues. The technique of solving this non-preemptive M/M/c model basically stems from Miller's approach, reader should refer to [20] for details of the algorithm.

The state space of the system is defined as follows :

1. Let $x_{i,j,k}$ be the state with i Type 1-SR waiting in the Type 1 infinite queue, j Type 2-SR waiting in the Type 2 infinite queue, k servers serving Type 1-SRs and $(c - k)$ servers serving Type 2-SRs, $i, j \geq 0$, $0 \leq k \leq c$. Then the blocks are defined as

$$H_{i,j} = \{x_{i,j,k} \mid 0 \leq k \leq c\}, \quad i, j \geq 0. \quad (4.1)$$

2. Let $x_{m,n}$ be the state with m server being in service, n Type 1-SR being serviced and no SR is waiting in any of Type 1 or Type 2 queue, $0 \leq m \leq (c - 1)$, $0 \leq n \leq \lfloor \frac{m}{b} \rfloor$,

Figure 4.3 The basic M/M/c two priorities queueing system

where $\lfloor \cdot \rfloor$ is the largest integer no greater than the argument. Then the blocks are defined as

$$H_m = \left\{ x_{m,n} \mid 0 \leq n \leq \left\lfloor \frac{m}{b} \right\rfloor \right\}, \quad 0 \leq m \leq (c-1). \quad (4.2)$$

Therefore, the state space is given as

$$S = \left(\bigcup_{m=0}^{c-1} H_m \right) \cup \left(\bigcup_{i \geq 0} \bigcup_{j \geq 0} H_{i,j} \right). \quad (4.3)$$

The matrices of transition rates between the blocks defined above are denoted as :

$$\begin{aligned} K_i &: H_i \rightarrow H_{i+1}, \quad 0 \leq i \leq c-1 \\ J_i &: H_i \rightarrow H_{i-1}, \quad 1 \leq i \leq c-1 \\ J_c &: H_{0,0} \rightarrow H_{c-1} \\ F_i &: H_i \rightarrow H_{i-b}, \quad b \leq i \leq c-1 \\ F_c &: H_{0,0} \rightarrow H_{c-b} \\ E_i &: H_i \rightarrow H_{i+b}, \quad 0 \leq i \leq c-b-1 \\ E_j &: H_j \rightarrow H_{0,0}, \quad c-b \leq j \leq c-1 \\ M_0 &: H_{i,j} \rightarrow H_{i-1,j}, \quad i \geq 1, j \geq 0 \\ M_i &: H_{0,j} \rightarrow H_{0,j-i}, \quad 1 \leq i \leq b, j \geq 1, j-i \geq 0 \\ X_i &: H_{0,j} \rightarrow H_{c-i+j}, \quad 2 \leq i \leq b, 0 \leq j \leq b-1, i+j \leq \text{rem}(c,b) \\ X'_i &: H_{0,j} \rightarrow H_{c-i+j}, \quad 2 \leq i \leq b, 0 \leq j \leq b-1, i+j > \text{rem}(c,b) \\ L_1 &: H_{i,j} \rightarrow H_{i+1,j}, \quad i, j \geq 0 \\ L_2 &: H_{i,j} \rightarrow H_{i,j+1}, \quad i, j \geq 0 \end{aligned} \quad (4.4)$$

* $\text{rem}(c,b)$ is the remainder of the division : c/b .

All the transitions shown previously correspond to arrivals and departures only and there is no transition within individual block. Hence, the submatrices of the transition rate matrix corresponding to transitions within a block are all diagonal and they are denoted as :

$$\begin{aligned} -D_i : H_i &\rightarrow H_i, 0 \leq i \leq c-1 \\ -D_c : H_{i,j} &\rightarrow H_{i,j}, i, j \geq 0 \end{aligned} \quad (4.5)$$

The state space with the transitions is shown in Fig.4.4.

The entry to each of the matrices mentioned above is given below :

$$\begin{aligned} K_i(p, q) &= \begin{cases} \lambda_2 & \text{if } q = p, i \neq c-1, 0 \leq p \leq \lfloor \frac{i}{b} \rfloor, 0 \leq q \leq \lfloor \frac{i+1}{b} \rfloor \\ \lambda_2 & \text{if } q = pb, i = c-1, 0 \leq p \leq \lfloor \frac{i}{b} \rfloor, 0 \leq q \leq c \\ 0 & \text{otherwise} \end{cases} \\ &\text{for } 0 \leq i \leq c-1 \\ J_i(p, q) &= \begin{cases} (i-bp)\mu_2 & \text{if } q = p, i \neq c, 0 \leq p \leq \lfloor \frac{i}{b} \rfloor, 0 \leq q \leq \lfloor \frac{i-1}{b} \rfloor \\ (c-p)\mu_2 & \text{if } q = \frac{p}{b}, i = c, 0 \leq p \leq c, 0 \leq q \leq \lfloor \frac{c-1}{b} \rfloor \\ p\mu_1/b & \text{if } q = \frac{p-1}{b}, i = c, 0 \leq p \leq c, 0 \leq q \leq \lfloor \frac{c-1}{b} \rfloor \\ 0 & \text{otherwise} \end{cases} \\ &\text{for } 1 \leq i \leq c \\ F_i(p, q) &= \begin{cases} p\mu_1 & \text{if } q = p-1, i \neq c, 0 \leq p \leq \lfloor \frac{i}{b} \rfloor, 0 \leq q \leq \lfloor \frac{i}{b} \rfloor - 1 \\ p\mu_1/b & \text{if } q = \frac{p}{b} - 1, b \neq 1, i = c, 0 \leq p \leq c, 0 \leq q \leq \lfloor \frac{c}{b} \rfloor - 1 \\ 0 & \text{otherwise} \end{cases} \\ &\text{for } b \leq i \leq c \\ E_i(p, q) &= \begin{cases} \lambda_1 & \text{if } q = p+1, i < c-b, 0 \leq p \leq \lfloor \frac{i}{b} \rfloor, 0 \leq q \leq \lfloor \frac{i}{b} \rfloor + 1 \\ \lambda_1 & \text{if } q = pb + c - i, i \geq c-b, 0 \leq p \leq \lfloor \frac{i}{b} \rfloor, 0 \leq q \leq c \\ 0 & \text{otherwise} \end{cases} \\ &\text{for } 0 \leq i \leq c-1 \end{aligned} \quad (4.6)$$

Figure 4.4 The partitioned state space (blocks) and transition scheme for Analysis 1

$$\begin{aligned}
M_0(p, q) &= \begin{cases} p\mu_1/b & \text{if } q = p, 0 \leq p, q \leq c \\ (c-p)\mu_2 & \text{if } q = p+1, \text{rem}(p, b) = 0, 0 \leq p, q \leq c \\ 0 & \text{otherwise} \end{cases} \\
M_i(p, q) &= \begin{cases} p\mu_1/b & \text{if } q = p-i, \text{rem}(q, b) = 0, 0 \leq p, q \leq c \\ (c-p)\mu_2 & \text{if } q = p, i = 1, \text{rem}(q, b) = 0, 0 \leq p, q \leq c \\ 0 & \text{otherwise} \end{cases} \\
&\text{for } 1 \leq i \leq b \\
X_i(p, q) &= \begin{cases} p\mu_1/b & \text{if } q = \frac{p-1}{b}, 0 \leq p \leq c, 0 \leq q \leq \lfloor \frac{c-1}{b} \rfloor \\ 0 & \text{otherwise} \end{cases} \\
&\text{for } 2 \leq i \leq b \\
X'_i(p, q) &= \begin{cases} p\mu_1/b & \text{if } q = \frac{p-1}{b}, 0 \leq p \leq c, 0 \leq q \leq \lfloor \frac{c-1}{b} \rfloor - 1 \\ 0 & \text{otherwise} \end{cases} \quad (4.7) \\
&\text{for } \max(2, \text{rem}(c, b) + 1) \leq i \leq b \\
L_i(p, q) &= \begin{cases} \lambda_i & \text{if } q = p, 0 \leq p, q \leq c \\ 0 & \text{otherwise} \end{cases} \\
&\text{for } i = 1, 2 \\
D_i(p, q) &= \begin{cases} \lambda_1 + \lambda_2 + p\mu_1 + (c-bp)\mu_2 & \text{if } q = p, 0 \leq p, q \leq \lfloor \frac{c}{b} \rfloor \\ 0 & \text{otherwise} \end{cases} \\
&\text{for } 0 \leq i \leq c-1 \\
D_c(p, q) &= \begin{cases} \lambda_1 + \lambda_2 + p\mu_1/b + (c-p)\mu_2 & \text{if } q = p, 0 \leq p, q \leq c \\ -(c-p)\mu_2 & \text{if } q = p+1, \text{rem}(p, b) \neq 0 \\ 0 & \text{otherwise} \end{cases}
\end{aligned}$$

The blocks $H_m, 0 \leq m \leq c-1$ and $H_{i,j}, i, j \geq 0$ are combined to form “super-blocks” and they are defined as follows :

$$I_{-1} = \bigcup_{i=0}^{c-1} H_i, \text{ and } I_i = \bigcup_{j=0}^{\infty} H_{i,j}, i \geq 0 \quad (4.8)$$

and the transition scheme is shown in Fig.4.5. The states in I_{-1} are exactly those corresponding to the existence of idle servers. The states in I_i , $i \geq 0$ are those with exactly i Type 1-SR awaiting service. The matrices of transition rates between and within these super-blocks are denoted as :

$$\begin{aligned}
 B_{-1,-1} &: I_{-1} \rightarrow I_{-1} \\
 B_{-1,0} &: I_{-1} \rightarrow I_0 \\
 B_{0,-1} &: I_0 \rightarrow I_{-1} \\
 B_{0,0} &: I_0 \rightarrow I_0 \\
 A_0 &: I_i \rightarrow I_{i+1}, i \geq 0 \\
 A_1 &: I_i \rightarrow I_i, i \geq 1 \\
 A_2 &: I_i \rightarrow I_{i-1}, i \geq 1
 \end{aligned} \tag{4.9}$$

The aforementioned matrices of transition rates are partitioned into submatrices corresponding to transitions between the sub-blocks of I_i , $i \geq -1$. The number of submatrices in one of these matrices varies according to the number of blocks in the corresponding super-blocks. The entry to each of the aforementioned matrices is described below :

$$B_{-1,-1}(p,q) = \begin{cases} -D_p & \text{if } q = p \\ K_p & \text{if } q = p+1, b \neq 1 \\ J_p & \text{if } q = p-1, b \neq 1 \\ E_p & \text{if } q = p+b, b \neq 1 \\ F_p & \text{if } q = p-b, b \neq 1 \\ K_p + E_p & \text{if } q = p+1, b = 1 \\ J_p + F_p & \text{if } q = p-1, b = 1 \\ 0 & \text{otherwise} \end{cases} \tag{4.10}$$

for $0 \leq p, q \leq c-1$

Figure 4.5 The partitioned state space (super-blocks) and transition scheme for Analysis 1

$$\begin{aligned}
B_{-1,0}(p, q) &= \begin{cases} E_p & \text{if } q = 0, c - b \leq p \leq c - 2 \\ E_{c-1} + K_{c-1} & \text{if } q = 0, p = c - 1 \end{cases} \\
&\quad \text{for } 0 \leq p \leq c - 1, 0 \leq q \leq \infty \\
B_{0,-1}(p, q) &= \begin{cases} J_c & \text{if } q = c - 1, p = 0 \\ Fc & \text{if } q = c - b, p = 0, b \neq 1 \\ X_i & \text{if } q = c - i + p, 2 \leq i \leq b, 0 \leq p \leq b - 1, \\ & q \geq [c - \text{rem}(c, b)], c - b + 1 \leq q \leq c - 1, \\ & \text{not}(i = b \text{ and } p = 0) \\ X'_i & \text{if } q = c - i + p, \max(2 \text{ rem}(c, b) + 1) \leq i \leq b, \\ & 0 \leq p \leq b - 1, q < [c - \text{rem}(c, b)], \\ & c - b + 1 \leq q \leq c - 1, \text{not}(i = b \text{ and } p = 0) \\ 0 & \text{otherwise} \end{cases} \\
&\quad \text{for } 0 \leq p \leq \infty, 0 \leq q \leq c - 1 \\
B_{0,0}(p, q) &= \begin{cases} -D_c & \text{if } q = p \\ L_2 & \text{if } q = p + 1 \\ M_i & \text{if } q = p - i, 1 \leq i \leq b \\ 0 & \text{otherwise} \end{cases} \\
&\quad \text{for } 0 \leq p, q \leq \infty \\
A_0(p, q) &= \begin{cases} L_1 & \text{if } q = p \\ 0 & \text{otherwise} \end{cases} \\
&\quad \text{for } 0 \leq p, q \leq \infty \\
A_1(p, q) &= \begin{cases} L_2 & \text{if } q = p + 1 \\ -D_c & \text{if } q = p \\ 0 & \text{otherwise} \end{cases} \\
&\quad \text{for } 0 \leq p, q \leq \infty \\
A_2(p, q) &= \begin{cases} M_0 & \text{if } q = p \\ 0 & \text{otherwise} \end{cases} \\
&\quad \text{for } 0 \leq p, q \leq \infty
\end{aligned} \tag{4.11}$$

Given the block structure and the notations adopted, the transition rate matrix for this M/M/c system has the form

$$Q = \begin{bmatrix} B_{-1,-1} & B_{-1,0} & 0 & 0 & 0 & \cdots \\ B_{0,-1} & B_{0,0} & A_0 & 0 & 0 & \cdots \\ 0 & A_2 & A_1 & A_0 & 0 & \cdots \\ 0 & 0 & A_2 & A_1 & A_0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \quad (4.12)$$

Invariant Measures

The process described by the above transition rate matrix is a special case of the "Markov Process of GI/M/1 Type" — Markov Process of Quasi-GI/M/1 Type with Complex Boundary Behaviour (QGM1-CB process). In fact, this process can also be regarded as the Quasi-Birth-and-Death (QBD) Process. As mentioned in chapter 2, the steady state probability vector of Q is of matrix geometric form where

$$\underline{\pi} = (\underline{\pi}_{-1}, \underline{\pi}_0, \underline{\pi}_0 R, \underline{\pi}_0 R^2, \underline{\pi}_0 R^3, \dots) \quad (4.13)$$

and satisfies

$$\underline{\pi} Q = \underline{0}, \quad \underline{\pi} \underline{e} = 1 \quad (4.14)$$

where $\underline{0}$ is a zero-row vector of infinite size and $\underline{e}$ is column vector of ones of appropriate size. The matrix R is called the "Rate matrix" of the QGM1-CB process and it is the minimal solution of

$$A_0 + R A_1 + R^2 A_2 = 0. \quad (4.15)$$

The (v, j) th element of R , R_{vj} ($v, j \geq 0$), is the expected time the system spent in $H_{i+1,j}$, ($i \geq 0$) before the first return to the set $\{H_{i,k}, k \geq 0\}$ for the process starting in $H_{i,v}$ [21,

20]. From the state space shown in Fig.4.4, it is clear that for the system to start in state $H_{i,v}$, the first return to the set $\{H_{i,k}, k \geq 0\}$ after visiting state $H_{i+1,j}$ can never reach the states $H_{i,w}, w < v$ due to the non-preemptive priority policy (i.e. Type 1-SR has to be served before any Type 2-SR when servers are available). Therefore, the matrix R has the following structure [21, 20] :

$$R = \begin{bmatrix} R_0 & R_1 & R_2 & \cdots \\ 0 & R_0 & R_1 & \cdots \\ 0 & 0 & R_0 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \quad (4.16)$$

where the submatrices R_i 's, $i \geq 0$ are matrices of dimension $(c+1) \times (c+1)$.

Substitution of Eqn.(4.16) into Eqn.(4.15) yields the following set of equations for solving the R_i 's, $i \geq 0$:

$$\begin{aligned} L_1 - R_0 D_c + R_0^2 M_0 &= 0 \\ R_{i-1} L_2 - R_i D_c + \sum_{j=0}^i R_j R_{i-j} M_0 &= 0, \quad i \geq 1 \end{aligned} \quad (4.17)$$

In order to solve the above set of equations for R_i 's, $i \geq 0$, it is first transformed to the equivalent set of equations describing the equivalent system of embedded Markov chain such that numerical iteration is possible. The embedded Markov chain is derived by observing the Markov process at times of state transition. Readers should refer to Appendix D for details.

M/G/1 Paradigm and Invariant Probability Vector

According to the theory of matrix-geometric invariant vector outlined in chapter 2, for the system which satisfies

$$\underline{\pi} \mathbf{Q} = \underline{0}, \quad \underline{\pi} \underline{e} = 1, \quad (4.18)$$

then the invariant probability vector $\underline{\pi}$ will have the form

$$\underline{\pi} = (\underline{\pi}_{-1}, \underline{\pi}_0, \underline{\pi}_1, \underline{\pi}_2, \dots) \quad (4.19)$$

where $\underline{\pi}_i$ corresponds to the probability vector on the super-block I_i , $i \geq -1$ and

$$\underline{\pi}_{i+1} = \underline{\pi}_i R, i \geq 0. \quad (4.20)$$

The boundary probability vector $\underline{\pi}_{-1}$, which corresponds to the probabilities of idle servers, and $\underline{\pi}_0$, which corresponds to the probabilities of system having busy servers with no Type 1-SR waiting and different number of Type 2-SR waiting, can be obtained first by solving the following subsystem :

$$(\underline{\pi}_{-1} \ \underline{\pi}_0) Q_{I_{-1} \cup I_0} = (\underline{0} \ \underline{0}) \quad (4.21)$$

where

$$Q_{I_{-1} \cup I_0} = \begin{bmatrix} B_{-1,-1} & B_{-1,0} \\ B_{0,-1} & B_{0,0} + RA_2 \end{bmatrix} \quad (4.22)$$

and

$$\begin{aligned} \underline{\pi}_{-1} &= (\pi_{-1,0}, \pi_{-1,1}, \dots, \pi_{-1,c-1}) \\ \underline{\pi}_0 &= (\pi_{0,0}, \pi_{0,1}, \pi_{0,2}, \dots) \end{aligned} \quad (4.23)$$

$$\underline{\pi}_{-1,i} = \text{Prob. vector on states in block } H_i, 0 \leq i \leq c-1$$

$$\underline{\pi}_{0,i} = \text{Prob. vector on states in block } H_{0,i}, i \geq 0$$

Note : Since $Q_{I_{-1} \cup I_0}$ is a matrix with negative diagonal entries, non-negative off-diagonal entries, and row sums equal to 0, it can be treated as a transition rate matrix for the Markov process on $I_{-1} \cup I_0$. Also, it is due to the special tridiagonal block structure of the matrix Q such that the all the transitions between the levels, which are the sets $\{H_{i,j} \mid j \geq 0\}, i > 0$, can be represented by the matrix RA_2 .

The matrix $Q_{I_{-1} \cup I_0}$ is of infinite size and most of the submatrices (partitions) in $B_{0,-1}$ and $B_{-1,0}$ are zero matrices. The solution to solving the boundary vectors is made possible due to the special structure of $(B_{0,0} + RA_2)$, which has the M/G/1 paradigm.

The matrix $(B_{0,0} + RA_2)$ has the following structure :

$$(B_{0,0} + RA_2) = \begin{bmatrix} C_1 & C_2 & C_3 & C_4 & \cdots \\ C_0 & C_1 & C_2 & C_3 & \cdots \\ 0 & C_0 & C_1 & C_2 & \cdots \\ 0 & 0 & C_0 & C_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \quad (4.24)$$

Each C_i 's, $i \geq 0$ is a square matrix of order $(b \times (c+1))$, formed by blocks of submatrices of size $(c+1) \times (c+1)$:

$$\begin{aligned} C_0 &= \begin{bmatrix} M_b & M_{b-1} & M_{b-2} & \cdots & M_1 \\ 0 & M_b & M_{b-1} & \cdots & M_2 \\ 0 & 0 & M_b & \cdots & M_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & M_b \end{bmatrix} \\ C_1 &= \begin{bmatrix} -D_c + R_0 M_0 & L_2 + R_1 M_0 & R_2 M_0 & \cdots & R_{b-1} M_0 \\ M_1 & -D_c + R_0 M_0 & L_2 + R_1 M_0 & \cdots & R_{b-2} M_0 \\ M_2 & M_1 & -D_c + R_0 M_0 & \cdots & R_{b-2} M_0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ M_{b-1} & M_{b-2} & M_{b-3} & \cdots & -D_c + R_0 M_0 \end{bmatrix} \\ C_2 &= \begin{bmatrix} R_b M_0 & R_{b+1} M_0 & R_{b+2} M_0 & \cdots & R_{2b-1} M_0 \\ R_{b-1} M_0 & R_b M_0 & R_{b+1} M_0 & \cdots & R_{2b-2} M_0 \\ R_{b-2} M_0 & R_{b-1} M_0 & R_b M_0 & \cdots & R_{2b-3} M_0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ L_2 + R_1 M_0 & R_2 M_0 & R_3 M_0 & \cdots & R_b M_0 \end{bmatrix} \end{aligned} \quad (4.25)$$

$$C_i = \begin{bmatrix} R_{(i-1)b}M_0 & R_{(i-1)b+1}M_0 & R_{(i-1)b+2}M_0 & \cdots & R_{ib-1}M_0 \\ R_{(i-1)b-1}M_0 & R_{(i-1)b}M_0 & R_{(i-1)b+1}M_0 & \cdots & R_{ib-2}M_0 \\ R_{(i-1)b-2}M_0 & R_{(i-1)b-1}M_0 & R_{(i-1)b}M_0 & \cdots & R_{ib-3}M_0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R_{(i-2)b+1}M_0 & R_{(i-2)b+2}M_0 & R_{(i-2)b+3}M_0 & \cdots & R_{(i-1)b}M_0 \end{bmatrix}, i \geq 3 \quad (4.26)$$

In order to solve the boundary vector $(\underline{\pi}_{-1} \ \underline{\pi}_0)$, it is necessary to convert the Markov process on $I_{-1} \cup I_0 = \left(\bigcup_{m=0}^{c-1} H_m \right) \cup \left(\bigcup_{j=0}^{\infty} H_{0,j} \right)$ to an embedded Markov chain on $I_{-1} \cup I_0$. It should also be noted that the matrix $(B_{0,0} + RA_2)$ is of infinite size (because of infinite queue size of Type 2-SR). Again, the special block structure of $(B_{0,0} + RA_2)$, which is of Hessenberg form, allow the transitions between the set of states in blocks $\{H_{0,j} \mid j \geq b\}$ be represented by the matrix $\hat{S}C_0$ such that the following boundary vector (with finite size) can be determined first :

$$(\underline{\pi}_{-1}, \underline{\pi}_{0,0}, \underline{\pi}_{0,1}, \dots, \underline{\pi}_{0,b-1}) \quad (4.27)$$

where

$$(\underline{\pi}_{-1}, \underline{\pi}_{0,0}, \underline{\pi}_{0,1}, \dots, \underline{\pi}_{0,b-1}) \tilde{\mathbf{Q}} = \underline{0} \quad (4.28)$$

The $(l, m)^{\text{th}}$ entry of the matrix $\hat{S}$, $0 \leq l, m \leq b \times (c+1) - 1$, is the expected time the system stays in $x_{0, (j+1)b + \lfloor \frac{m}{b(c+1)} \rfloor, \text{rem}(m, b(c+1))}$, $j \geq 0$ starting from state $x_{0, j b + \lfloor \frac{l}{b(c+1)} \rfloor, \text{rem}(l, b(c+1))}$ before returning to the states in $\bigcup_{n=0}^{b-1} H_{0, j b + n}$. $\hat{S}$ is the first passage time distribution matrix. The $\tilde{\mathbf{Q}}$ matrix, which describes the transitions between the set of states in blocks $\left(\bigcup_{m=0}^{c-1} H_m \right) \cup \left(\bigcup_{j=0}^{b-1} H_{0,j} \right)$, has the following structure :

vectors for the various states, are given in Appendix D.

4.4 Analysis 2 — Non-Preemptive Priority M/M/c Model : Type 2–SR has priority over Type 1–SR

In this model Type 2–SR has absolute non-preemptive priority over Type 1–SR. When any server is occupied by Type 1–SR, Type 2–SR cannot preempt. When any server is released upon the departure of any SR, Type 2–SR has priority over Type 1–SR in seizing the servers for service. The approach of the analysis is similar to Analysis 1, hence the analysis will be described briefly. The entries to the various matrices shall be referred to Appendix E.

The major difference in the approach of solving the system between the previous analysis (Analysis 1) and the current analysis are that instead of combining all the states describing Type 2–SRs waiting in the queue to form the super-blocks, the super-blocks are formed by combining all the states describing Type 1–SRs waiting in the queue.

Using the same definitions for the system states and blocks as in Analysis 1, the state space is again given as

$$S = \left(\bigcup_{m=0}^{c-1} H_m \right) \cup \left(\bigcup_{i \geq 0} \bigcup_{j \geq 0} H_{i,j} \right). \quad (4.30)$$

The matrices of transition rates between and within the blocks, such as D_i , K_i , J_i , X_i , $i \geq 0$ etc., and the entries to each of these matrices are given in Appendix E. The state space with the transitions given is shown in Fig.4.6. The blocks H_m , $0 \leq m \leq c-1$ and $H_{i,j}$, $i, j \geq 0$ are combined to form “super-blocks” and they are defined as follows :

$$I_{-1} = \bigcup_{i=0}^{c-1} H_i, \text{ and } I_j = \bigcup_{i=0}^{\infty} H_{i,j}, j \geq 0 \quad (4.31)$$

and the transition scheme is shown in Fig.4.7. The states in I_{-1} are exactly those corresponding to the existence of idle servers. The states in I_j , $i \geq 0$ are those with

Figure 4.6 The partitioned state space (blocks) and transition scheme for Analysis 2

Figure 4.7 The partitioned state space (super-blocks) and transition scheme for Analysis 2.

exactly i Type 2-SR awaiting service. The matrices of transition rates between and within these super-blocks are denoted as :

$$\begin{aligned}
 B_{i,-1} &: I_i \rightarrow I_{-1}, \quad -1 \leq i \leq b-1 \\
 B_{i,0} &: I_i \rightarrow I_0, \quad -1 \leq i \leq b-1 \\
 A_j &: I_i \rightarrow I_{i-j+1}, \quad i \geq 0, \quad 0 \leq j \leq b+1, \quad i-j \geq 0 \\
 A_{b+1} &: I_b \rightarrow I_0
 \end{aligned} \tag{4.32}$$

The aforementioned matrices of transition rates are partitioned into submatrices corresponding to the transitions between the sub-blocks of I_i , $i \geq -1$. The entry (the sub-matrix) to each of the matrices is provided in Appendix E. Given the block structure and the notations adopted, it can be shown clearly from Fig.4.7 that the transition rate matrix for this M/M/c system has the form

$$Q = \begin{bmatrix} B_{-1,-1} & B_{-1,0} & 0 & 0 & 0 & \cdots \\ B_{0,-1} & B_{0,0} & A_0 & 0 & 0 & \cdots \\ B_{1,-1} & B_{1,0} & A_1 & A_0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \\ B_{b-1,-1} & B_{b-1,0} & A_{b-1} & A_{b-2} & A_{b-3} & \cdots \\ 0 & A_{b+1} & A_b & A_{b-1} & A_{b-2} & \cdots \\ 0 & 0 & A_{b+1} & A_b & A_{b-1} & \cdots \\ 0 & 0 & 0 & A_{b+1} & A_b & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{4.33}$$

Invariant Measures

The process described by the above transition rate matrix is again a Markov Process of Quasi-GI/M/1 Type with Complex Boundary Behaviour (QGM1-CB process). As

mentioned in chapter 2, the steady state probability vector of $\mathbf{Q}$ is of matrix geometric form where

$$\underline{\pi} = (\pi_{-1}, \pi_0, \pi_1, \pi_2, \pi_3, \dots) = (\pi_{-1}, \pi_0, \pi_0 R, \pi_0 R^2, \pi_0 R^3, \dots) \quad (4.34)$$

and satisfies

$$\underline{\pi} \mathbf{Q} = \underline{0}, \quad \underline{\pi} \underline{e} = 1 \quad (4.35)$$

where $\underline{0}$ is a zero-row vector of infinite size and $\underline{e}$ is column vector of ones of appropriate size. The rate matrix R of this QGM1-CB process is the minimal solution of

$$A_0 + RA_1 + R^2 A_2 + \dots + R^{b+1} A_{b+1} = 0. \quad (4.36)$$

The definition of R is similar to that given in Analysis 1 and the matrix R again has the same Toeplitz structure [21, 20] :

$$R = \begin{bmatrix} R_0 & R_1 & R_2 & \dots \\ 0 & R_0 & R_1 & \dots \\ 0 & 0 & R_0 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \quad (4.37)$$

where the submatrices R_i 's, $i \geq 0$ are matrices of dimension $(c+1) \times (c+1)$.

Similar to Analysis 1, substitution of Eqn.(4.37) into Eqn.(4.36) will yield one set of simultaneous equations, which is used to solve the R_i 's, $i \geq 0$. It should be noted that slightly complicated iteration procedure is involved in solving R_i 's due to the $(b+1)$ th order matrix equation given by Eqn.(4.36). Readers should refer to Appendix E for the numerical procedures.

M/G/1 Paradigm and Invariant Probability Vector

Similar to Analysis 1, the invariant probability vector of process $\mathbf{Q}$, $\underline{\pi}$, is a matrix-geometric invariant vector and have the form

$$\underline{\pi} = (\pi_{-1}, \pi_0, \pi_1, \pi_2, \dots) \quad (4.38)$$

which satisfies

$$\underline{\pi} \mathbf{Q} = \underline{0}, \quad \underline{\pi} \underline{e} = 1, \quad (4.39)$$

and $\underline{\pi}_i$ corresponds to the probability vector on the super-block I_i , $i \geq -1$ with

$$\underline{\pi}_{i+1} = \underline{\pi}_i R, \quad i \geq 0 \quad \text{or} \quad \underline{\pi}_{i+1} = \underline{\pi}_0 R^{i+1}, \quad i \geq 0. \quad (4.40)$$

The boundary probability vector $\underline{\pi}_{-1}$, which corresponds to the probabilities of idle servers, and $\underline{\pi}_0$, which corresponds to the probabilities of system having busy servers with no Type 2-SR waiting and different number of Type 1-SR waiting, can be obtained first by solving the following subsystem :

$$(\underline{\pi}_{-1} \quad \underline{\pi}_0) Q_{I_{-1} \cup I_0} = (\underline{0} \quad \underline{0}) \quad (4.41)$$

where

$$Q_{I_{-1} \cup I_0} = \begin{bmatrix} B_{-1,-1} & B_{-1,0} \\ \sum_{i=0}^{b-1} R^i B_{i,-1} & \sum_{i=0}^{b-1} R^i B_{i,0} + R^b A_{b+1} \end{bmatrix} \quad (4.42)$$

and

$$\underline{\pi}_{-1} = (\pi_{-1,0}, \pi_{-1,1}, \dots, \pi_{-1,c-1})$$

$$\underline{\pi}_0 = (\pi_{0,0}, \pi_{1,0}, \pi_{2,0}, \dots)$$

(4.43)

$$\underline{\pi}_{-1,i} = \text{Prob. vector on states in block } H_i, \quad 0 \leq i \leq c-1$$

$$\underline{\pi}_{i,0} = \text{Prob. vector on states in block } H_{i,0}, \quad i \geq 0$$

Note : Similar as in Analysis 1, $Q_{I_{-1} \cup I_0}$ can be treated as a transition rate matrix for the Markov process on $I_{-1} \cup I_0$ and the matrix $\left(\sum_{i=0}^{b-1} R^i B_{i,0} + R^b A_{b+1} \right)$ also has the M/G/1 paradigm. Following the same approach as outlined in Analysis 1, the boundary vector $(\underline{\pi}_{-1} \quad \underline{\pi}_0)$ on the state space $I_{-1} \cup I_0 = \left(\bigcup_{m=0}^{c-1} H_m \right) \cup \left(\bigcup_{i=0}^{\infty} H_{i,0} \right)$ can first be determined by computing

$$(\underline{\pi}_{-1}, \underline{\pi}_{0,0}). \quad (4.44)$$

Figure 4.8 The preemptive priority with boundary M/M/c queueing system

Then the matrices of “first passage time distribution” are used with the vector $(\pi_{-1}, \pi_{0,0})$ to compute the probability vectors for states $H_{j,0}$, $j \geq 1$ ($= \pi_{1,0}, \pi_{2,0}, \pi_{3,0}, \dots, \pi_{\infty,0}$). With the rate matrix R , the probability vectors for states $H_{i,j}$, $i \geq 0$, $j \geq 1$ ($= \pi_1, \pi_2, \dots, \pi_{\infty}$) can be computed recursively. For detailed procedures of the analysis, one is referred to Appendix E.

4.5 Analysis 3 — Preemptive Priority with Boundary M/M/c Model

In this model, Type 1-SRs have a preemptive priority over Type 2-SR in c_1 ($\leq c$) servers only and that Type 2-SRs have a preemptive priority over Type 1-SR in the remaining c_2 ($= c - c_1$) servers. The model is illustrated in Fig.4.8.

The state space of the system is

$$S = \{(n_1, n_2), n_1, n_2 \geq 0\} \quad (4.45)$$

where n_i ($i = 1, 2$) is the number of Type i -SRs in the system (see Fig.4.9). Since it is a preemptive priority (PP) system, the number of Type i -SRs ($i = 1, 2$) *waiting* in the respective infinite size queues are readily known from the number of Type i -SRs in the system.

Denote the following :

$$p_{i,j} = \text{Prob. } [n_1 = i, n_2 = j] \quad (4.46)$$

By combining the set of states $\{(n_1 = i, n_2 = j) \mid j \geq 0\}$ to become a super-state, H_i ($i \geq 0$), the corresponding probability vector for that super-state is

$$\underline{p}_i = (p_{i,0}, p_{i,1}, p_{i,2}, \dots, p_{i,\infty}). \quad (4.47)$$

Then the system can be modelled as a Markov process with the transition rate matrix $\mathbf{Q}$ given by

$$\mathbf{Q} = \begin{bmatrix} B_0 & D_0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ C_1 & B_1 & D_1 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & C_2 & B_2 & D_2 & 0 & 0 & 0 & 0 & \dots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & C_r & B_r & D_r & 0 & 0 & \dots \\ 0 & 0 & \dots & 0 & A_2 & A_1 & A_0 & 0 & \dots \\ 0 & 0 & \dots & 0 & 0 & A_2 & A_1 & A_0 & \dots \\ \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \quad (4.48)$$

and its invariant probability vector $\underline{p}$ given by

$$\underline{p} = (\underline{p}_0, \underline{p}_1, \underline{p}_2, \dots) \quad (4.49)$$

which satisfies

$$\underline{p}\mathbf{Q} = \underline{0} \text{ and } \underline{p}\underline{e} = 1. \quad (4.50)$$

Figure 4.9 The state space and transition scheme for PP system (Analysis 3)

Note : $r = \lceil \frac{c}{b} \rceil - 1$ where $\lceil \cdot \rceil$ is the smallest integer not less than the argument, and $\underline{0}$ is an infinite size zero vector and $\underline{e}$ is an infinite size unit vector.

Each of the submatrices A_i 's, B_i 's, C_i 's is 2 dimensional matrix of infinite size and the elements are given as follows :

$A_0 = \lambda_1 I$, where I is an identity matrix

$$\begin{aligned}
 A_1(i, j) &= \begin{cases} \lambda_2 & \text{if } j = i + 1 \\ -(\lambda + \zeta\mu_1 + \min(j, c - b\zeta)\mu_2) & \text{if } j = i \\ \min(j, c - b\zeta)\mu_2 & \text{if } j = i - 1 \\ 0 & \text{otherwise} \end{cases} \\
 &\text{for } 0 \leq i, j \leq \infty \\
 A_2(i, j) &= \begin{cases} \zeta\mu_1 & \text{if } j = i \\ 0 & \text{otherwise} \end{cases} \\
 &\text{for } 0 \leq i, j \leq \infty \\
 B_m(i, j) &= \begin{cases} \lambda_2 & \text{if } j = i + 1 \\ -(\lambda + \min(\zeta, m)\mu_1 + \min(j, c - b(\min(\zeta, m)))\mu_2) & \text{if } j = i \\ \min(i, c - b(\min(\zeta, m)))\mu_2 & \text{if } j = i - 1 \\ 0 & \text{otherwise} \end{cases} \\
 &\text{for } 0 \leq m \leq r, 0 \leq i, j \leq \infty \\
 C_m(i, j) &= \begin{cases} \min(\zeta, m)\mu_1 & \text{if } j = i \\ 0 & \text{otherwise} \end{cases} \\
 &\text{for } 1 \leq m \leq r, 0 \leq i, j \leq \infty \\
 D_m &= A_0, 0 \leq m \leq r
 \end{aligned} \tag{4.51}$$

where $\zeta = \frac{c - \min(j, c_2)}{b}$ and $\lambda = \lambda_1 + \lambda_2$.

The Markov process shown is the QBD process outlined in Chapter 2. Following Neuts' method [1], the stationary probability vector for this system (which has a stable queue), $\underline{p}$, partitioned into probability vectors which correspond to each super-state, is given as

$$\underline{p} = (\underline{p}_0, \underline{p}_1, \underline{p}_2, \dots, \underline{p}_{r-1}, \underline{p}_r, \underline{p}_{r+1}, \dots) \tag{4.52}$$

where

$$\underline{p}_i = \underline{p}_r R^{i-r}, \text{ for } i \geq r \quad (4.53)$$

and R (the rate matrix) is the minimal non-negative solution to the equation

$$R^2 A_2 + R A_1 + A_0 = 0. \quad (4.54)$$

The boundary probability vectors $\underline{p}_0, \underline{p}_1, \underline{p}_2, \dots, \underline{p}_{r-1}, \underline{p}_r$ satisfy either one of the following sets of equations :

$$\begin{aligned} \text{for } r > 0 : & \begin{cases} \underline{p}_0 B_0 + \underline{p}_1 C_1 = 0 \\ \underline{p}_{i-1} D_{i-1} + \underline{p}_i B_i + \underline{p}_{i+1} C_{i+1} = 0, & \text{for } 1 \leq i \leq r-1 \\ \underline{p}_{r-1} D_{r-1} + \underline{p}_r (B_r + R A_2) = 0 \\ \sum_{v=0}^{r-1} \underline{p}_v \underline{e} + \underline{p}_r (I - R)^{-1} \underline{e} = 1 \end{cases} \\ \text{for } r = 0 : & \begin{cases} \underline{p}_0 B_0 + \underline{p}_1 A_1 = 0 \\ \underline{p}_0 A_0 + \underline{p}_1 (A_1 + R A_2) = 0 \\ \underline{p}_0 (I - R)^{-1} \underline{e} = 1 \end{cases} \end{aligned} \quad (4.55)$$

Solving the set of equations yield $\underline{p}_i$ ($0 \leq i \leq r$). Note that the last equation in each of the above sets is the normalization condition and $\underline{e}$ is unit column vector.

Note 1 : If a Type 1-SR is allowed to be served only when sufficient number of servers (i.e. b) are available, the analysis is still applicable with the modification $\zeta = \left\lfloor \frac{c - \min(j, c_2)}{b} \right\rfloor$ where $\lfloor \cdot \rfloor$ is the largest integer not greater than the argument.

Note 2 : Due to the nature of the PP policy and the way the super-state is defined, it is possible for the system to start in state $(n_1 = i, n_2 = v), v > 0$, and then on the first return to the set $\{(n_1 = i, n_2 = k), k \geq 0\}$ after visiting state $(n_1 = i + 1, n_2 = j), j \geq 0$, it reaches the state $(n_1 = i, n_2 = w), w < v$. Hence the nice upper-triangular block

structure of R , which appears in the Analysis 1 and Analysis 2, does not exist and the structure of R cannot easily be determined.

Note 3 : Since the analysis cannot make use of the Toeplitz structure of R and hence the infinite size matrices such as A_i , B_i , C_i etc. have to be arbitrarily fixed such that numerical analysis is possible. The size of the matrices should be chosen large enough to minimize the inaccuracy incurred in the analysis. The infinite size of the matrices corresponds to the infinite Type 2 system size. Two suggestions for approximating the Type 2 system size are given in Appendix F. A brief discussion on the actual size of the matrices and the computation times required are given in section 4.6.

Note 4 : For computational efficiency, the rate matrix R and the probability vectors of the corresponding embedded Markov chain are obtained first. The embedded Markov chain is derived by observing the Markov process at the times of state transition. The probability vectors, $\underline{p}_i$ ($0 \leq i \leq r$), are then obtained by appropriately transforming those of the embedded Markov chain, and the rest, $\underline{p}_i$ ($i > r$), are obtained using the rate matrix R as given in Eqn.(4.53).

The detailed procedures for solving the probability vectors of this PP system are given in Appendix F.

4.6 Numerical Examples and Discussions

In each set of the numerical examples shown in this section, the parameters λ_1 , μ_1 , μ_2 , b , c and c_2 (for Analysis 3) are fixed and the system utilizations are varied by choosing different values of λ_2 . The system utilization, ρ , is given by $\rho = \frac{b\lambda_1}{c\mu_1} + \frac{\lambda_2}{c\mu_2}$. For the analysis of the system performance where PP policy is employed (Analysis 3), the parameter c_2 is the number of servers where Type 2–SR has preemptive priority over Type 1–SR. In the comparison of performance between NPP and PP policies, the

definition of the parameter c_2 is relaxed for convenience to mean the number of servers where Type 2–SR has “priority” over Type 1–SR. Therefore, priority policy in Analysis 1 is a special case of non-preemptive priority policy where $c_2 = 0$, and Analysis 2 is again a special case where $c_2 = c$.

A numerical example of Analysis 1 showing the queue length distributions for both Type 1–SR and Type 2–SR under various system utilizations is given in Fig.4.10. The parameter values chosen are : $\lambda_1=1.0$, $\mu_1=1.0$, $\mu_2=0.1$, $b=6$, and $c=20$. The system utilizations, which range from 0.4 to 0.9, are varied by increasing λ_2 . Since Type 1–SR has absolute non-preemptive priority over Type 2–SR and its average service time is 10 times less than that of Type-2 SR, increasing Type 2–SR's arrival rate does not affect Type 1–SR's queueing performance as significant as of Type 2–SR.

Fig.4.11 shows an example of Analysis 2 with $\lambda_1=0.1$, $\mu_1=0.1$, $\mu_2=1.0$, $b=4$, and $c=10$. In this case, Type 2–SR's arrival is more frequent than Type 1–SR but its average service time is 10 times less than that of Type 1–SR. Although Type 2–SRs have absolute non-preemptive priority over Type 1–SRs in accessing the servers, Type 1–SR has long service time on average and uses up to a maximum of 4 out of 10 servers for service, hence increasing Type 2–SR's arrival rate causes both Type 1 and Type 2–SRs' queue size to increase quite significantly.

Queue length distribution can also be obtained for the PP system with the matrix-analytic methodology. In the following examples, instead of presenting the queue length distributions for the system having PP policy, the mean waiting times for Type 1–SR and Type 2–SR under different values of c_2 with different set of service rates are plotted against the system utilization in Fig.4.12 and Fig.4.13. Examples of queue length distributions for PP system are given later. In Fig.4.12, the common parameter values are chosen as : $\lambda_1=0.1$, $\mu_1=0.1$, $\mu_2=1.0$, $b=3$, and $c=20$. In Fig.4.13, the set of parameter values are : $\lambda_1=1.0$, $\mu_1=1.0$, $\mu_2=0.1$, $b=6$, and $c=20$. The dotted line

Figure 4.10 Queue length distribution for (a) Type 1-SR and (b) Type 2-SR under various utilizations with Type 1-SR has non-preemptive priority over Type 2-SR ($b=6$, $c=20$, $\lambda_1=1.0$, $\mu_1=1.0$, $\mu_2=0.1$)

Figure 4.11 Queue length distribution for (a) Type 1-SR and (b) Type 2-SR under various utilizations with Type 2-SR has non-preemptive priority over Type 1-SR ($b=4$, $c=10$, $\lambda_1=0.1$, $\mu_1=0.1$, $\mu_2=1.0$)

represents the waiting time when Type 1-SR are not served until b servers are available, while the solid line represents the waiting time when Type 1-SR begins service even if less than b servers are available. As mentioned in the previous sections, the service rate is reduced accordingly when less than b servers serve a Type 1-SR. As expected, the waiting time of Type 1-SR decreases considerably when Type 1-SR are allowed to use less than b servers especially when the system utilization and c_2 are large. As for the waiting time of Type 2-SR in this example is concerned, it is rather insensitive to whether Type 1-SR is allowed to begin service when less than b servers are available or it must wait to enter service until b servers are available, even when c_2 is small.

In order to find the sensitivity of the waiting times to the changes in c_2 more explicitly, an example is shown in Fig.4.14 where $b=4$, $c=20$, $\lambda_1/\lambda_2=10$, $\mu_1=1.0$, $\mu_2=0.1$ and c_2 is varied for three values of the system utilization, 0.25, 0.60 and 0.85. The waiting times do not change significantly with c_2 under low and medium utilizations; however, at high utilizations the changes in waiting time is considerable near $c_2 = \frac{c}{1+b(\lambda_1\mu_2)/(\lambda_2\mu_1)}$ ($= 4$ in the example). Note that the result for system utilization $\rho = 0.25$ are very close to the x-axis.

The comparisons between the queueing performance for systems employing PP policy and NPP policy are shown in Fig.4.15 to Fig.4.17. Three different sets of parameter values are chosen for each of the examples. In Fig.4.15, the mean waiting times for both Type 1-SR and Type 2-SR under the two different types of priority policies are compared with parameter values $\lambda_1=1.0$, $\mu_1=\mu_2=1.0$, $b=3$ and $c=10$. For the case of $c_2 = 3$ and $c_2 = 6$ of the NPP policy, the result is obtained from simulations. For this set of parameter values, as expected, Type 1-SR favors the preemptive-priority policy when c_2 is small and non-preemptive-priority policy when c_2 is large, and vice versa for Type 2-SR. However, the difference in mean waiting times between PP and NPP policy is not significant for the various values of c_2 .

Figure 4.12 Mean waiting time for (a) Type 1-SR and (b) Type 2-SR vs. system utilization (ρ) with system employs PP policy ($b=3$, $c=20$, $\lambda_1=0.1$, $\mu_1=0.1$, $\mu_2=1.0$)

Figure 4.13 Mean waiting time for (a) Type 1-SR and (b) Type 2-SR vs. system utilization (ρ) with system employs PP policy ($b=6$, $c=26$, $\lambda_1=1.0$, $\mu_1=1.0$, $\mu_2=0.1$)

Figure 4.14 Effect of NB group (c_2) on mean waiting time for (a) Type 1-SR and (b) Type 2-SR with PP policy ($b=4$, $c=20$, $\lambda_1/\lambda_2=10$, $\mu_1=1.0$, $\mu_2=0.1$)

Fig.4.16 and Fig.4.17 show the significant differences in queueing performance between system employing PP and NPP policies. In both examples, the system utilization is 0.7. In Fig.4.16, $\lambda_1=1.0$, $\mu_1=1.0$, $\lambda_2=0.8$, $\mu_2=0.1$, $b=6$, $c=20$ and $c_2=0$. Obviously, Type 2-SR would favor NPP policy and vice versa for Type 1-SR since when $c_2=0$ and if priority policy is PF, Type 1-SR can always preempt Type 2-SR over all the servers. For this set of parameter values, the choice between NPP and PP policy is subjected to whether Type 1-SRs or Type 2-SRs have stringent performance requirements over the other. However, for the set of parameter values used in Fig.4.17 where $\lambda_1=0.1$, $\mu_1=0.1$, $\lambda_2=3.0$, $\mu_2=1.0$, $b=4$, $c=10$, and $c_2=10$, PP policy is definitely superior to NPP policy.

Different combinations of b , c , λ_1 , μ_1 , λ_2 , and μ_2 have also been tested. It is found that the numerical solutions are sometimes difficult to obtain under certain combinations of parameter values. For example, in Analysis 1 where Type 1-SR has non-preemptive priority over Type 2-SR, if Type 2-SR's arrival rate is high (e.g. 10 times larger than that of Type 1-SR's) and its service rate is low (e.g. 10 times less than that of Type 1-SR's) and/or Type 1-SR service rate is low, the convergence rate of the various matrices such as G , R etc. (see Appendix D) is low and hence very long iterations are required.

Using a Sun 3/60 20 Mhz workstation with a share of 30% cpu and the pro-Matlab package, average computation time for such a set of parameter values aforementioned is about 8 to 12 hours. Similar experience is encountered in Analysis 2 or 3 when some combinations of parameter values chosen are skewed, i.e. the ratio between λ_1 and λ_2 , and/or μ_1 and μ_2 is large (> 10) or small (< 0.1). In the case of PP policy, the average computation time is about 4 to 6 hours. When the parameter values are not skewed, the average computation time for one set of parameter values using the NPP policy (Analysis 1 and 2) is approximately 2 to 4 hours, where for PP policy the computation time ranges from 15 minutes to 1 hour under similar cpu loading. More reliable simulation results can be obtained where simulation time for the NPP and PP

Figure 4.15 Comparison of mean waiting time for (a) Type 1-SR and (b) Type 2-SR under various utilizations between NPP and PF policy ($b=3$, $c=10$, $\lambda_1=1.0$, $\mu_1=1.0$, $\mu_2=1.0$)

Figure 4.16 Comparison of queue length distribution for (a) Type 1-SR and (b) Type 2-SR between NPP and PP policy ($b=6$, $c=20$, $c_2=0$, $\lambda_1=1.0$, $\mu_1=1.0$, $\lambda_2=0.8$, $\mu_2=0.1$, $\rho=0.7$)

Figure 4.17 Comparison of queue length distribution for (a) Type 1-SR and (b) Type 2-SR between NPP and PP policy ($b=4$, $c=10$, $c_2=c$, $\lambda_1=0.1$, $\mu_1=0.1$, $\lambda_2=3.0$, $\mu_2=1.0$, $\rho=0.7$)

policies is about 20 to 26 hours on average in the case of skewed parameter values, and it is about 10 to 14 hours in the case of less skewed parameter values.

In the numerical implementation, the size of the matrix R and other relevant matrices in the analysis of the FP policy are determined (or fixed) such that numerical implementation is made possible. In the case of skewed parameter values and high system utilization, the average size of the matrix is approximately 70×70 , while in the case of less skewed parameter values, the average size of the matrix is approximately 10×10 . With these values, the accuracy of the analysis can still be maintained.

4.7 Conclusions

In the study, two types of service-requests (Type 1-SR and Type 2-SR) acquiring access to the ISDN through an access controller are considered where Type 1-SR requires a maximum of b BBUs and Type 2-SR requires only 1 BBU out of c BBUs for service. The basic system can be considered as a $M/M/c$ queueing system. The queueing performance for both Type 1-SR and Type 2-SR under three different types of access control strategies is studied. The three strategies are : 1) Type 1-SR has non-preemptive priority over Type 2-SR (Analysis 1); 2) Type 2-SR has non-preemptive priority over Type 1-SR (Analysis 2); and 3) Type 2-SR has preemptive priority over Type 1-SR in c_2 servers while Type 1-SR has preemptive priority over Type 2-SR in the rest of the servers, $c - c_2$ (Analysis 3). Common to all three strategies are that Type 1-SR can enter service with the service time properly adjusted in case that the number of servers available is less than b , and both Type 1-SR and Type 2-SR are queueable.

The state spaces for the two NPP strategies are complicated. Numerical solutions are made possible due to special upper-triangular block structure of the rate matrix R where individual partition in R can be computed recursively. Numerical examples for the three

types of access control strategies with different combinations of parameter values are given. It is shown that the choice between PP policy and NPP policy should be dependent on the set of parameter values chosen in order to achieve a better overall performance for both Type 1-SR and Type 2-SR. Despite the fact that the analysis for the PP system is comparatively easy, numerical results indicate that PP system is not always a good approximation to the NPP system. Attempt to approximate the queueing behaviour of NPP system with a PP queueing system shall be accompanied by a careful selection of parameter values. Numerical difficulties are also encountered sometimes when the set of parameter values is skewed. They are mainly due to the slow convergence in the iteration procedures for solving certain matrices and/or the boundary probability vectors. In the case of PP policy, the analysis is capable of solving an arbitrary c_2 , $0 \leq c_2 \leq c$. The analysis for the case of NPP policy is only capable of solving for $c_2=0$ or $c_2=c$. Since it is the special structure of R which makes the analysis of non-preemptive system possible, further studies on the rate matrix R are required such that the analysis for NPP system may become feasible with $0 < c_2 < c$.

Note : The analysis presented in this chapter is mainly applied to the call admission level at the access point of the Broadband-ISDN. In the burst level analysis of the Broadband-ISDN, which may incorporate the concepts of virtual connection and virtual path, similar stochastic model (M/M/c two-priority model) can be derived. Since a single virtual path has fixed amount of capacity and is shared with different number of virtual connections, the virtual path capacity can be subdivided into c units with the burst in the individual virtual connection requiring a portion of the virtual path capacity to transmit. The virtual path capacity is equivalent to the access link capacity and the burst is equivalent to the service-request in this study. This analysis is also applicable in that situation.

Chapter 5

A Study on the Markov Modulated Characterization of Packetized Videophone Traffic and related Statistical Multiplexer Performance with the application of Access Control Strategy

5.1 Introduction

A major traffic component on Integrated Services Digital Network will be the digital video. Applications of such type of service impose very large bandwidth requirements on communications network. Variable-bit-rate compression algorithms have been proposed to reduce the bandwidth by transmitting at a higher rate during high-activity (motion) scenes and at a lower rate when there is less motion. With variable-bit-rate video (as well as superposition of other types of services), the input process to the statistical multiplexer is often complex and can possess correlations in the number of packet arrivals in adjacent time intervals. Related queueing performance would also be affected significantly.

Exact analysis for systems to which the input process is correlated is often intractable. In addition, the exact modelling of the correlated video sources (which can be used in mathematical analysis) is not well known. Recently, Maglaris *et al.* have studied

Figure 5.1 Videophone Source

the variable-bit-rate videophone source (see Fig.5.1 for illustration) and they have proposed two models for packet switching of the videophone sources [22]. The first model is a Continuous-State First-order Auto-Regressive Markov model, which is rather complicated for mathematical analysis and hence it is used by Maglaris *et al.* for simulation purposes only. The second (analytic) model is the Discrete-State Continuous-Time Markov Process (DS-CTMP), but its applicability is limited due to the large number of states it requires in order to provide a good approximation for the correlated process (see Fig.5.2).

A method proposed in [23] uses a correlated 2-state Markov Modulated Poisson Process (MMPP) to approximate the arrival process of a superposition of packetized voice sources (On-Off source models). The method provides a fairly good approximation within a certain range of parameter values. If the 2-state MMPP model is a good approximation to the videophone source model (which is shown in [22] as an aggregation of many independent identical On-Off sources for each individual video source), perhaps

Figure 5.2 Approximation for Videophone source output rate

the performance analysis of the statistical multiplexer with correlated arrival process can be made feasible since the number of states required to provide a good approximation to the aggregated correlated arrival process is dramatically reduced. Although the MMPP model is an approximation to the discrete-state continuous-time Markov model (which itself is an approximation to the actual videophone sources), it at least provides a mathematically tractable model for correlated sources. Since the dimensionality of the model is kept small and especially when there is no well-known model for the video sources, MMPP can be a potential candidate for correlated source modelling.

Another important issue which arises very often in the ISDN environment is that there are services such as voice and video which usually have stringent requirements on delay and cell loss. Access control strategy (priority access) should be exercised at the cell transport level of the multiplexer to guarantee the grade of service. It is often difficult to obtain an exact analysis on the queueing behaviour of the low priority service when the high priority services possess correlation in the arrival process. Simulations are typically used in studying the performance issues for this class of problems, however, it is expensive and is not capable of obtaining low probability tails unless long simulation time is exploitable. A good approximation method, which is mathematically tractable and require less computing time, is hence desirable.

In this chapter, a preliminary study on the suitability of a 2-state Markov Modulated Poisson Process to represent videophone sources is presented. Also, an approximation on the queueing behaviour of the low priority packets in the presence of the high priority correlated packets (videophone packets) is proposed and analyzed via the matrix-analytic approach. The outline of this chapter is as follows: Section 5.2 introduces the discrete-state continuous-time Markov model as described in [22] and presents the 2-state MMPP model as a model for aggregated videophone sources. The exact queueing analysis of high priority packets and the approximation of the queueing behaviour of low

priority packets in a multi-servers system (multiple output channels at the multiplexer) is presented in section 5.3 . Numerical examples showing the suitability of the MMPP model and the queueing performance of the two priorities queueing system are given in section 5.4. Conclusions for the study are followed in section 5.5.

5.2 The Videophone Source Models

Two analytic models for videophone sources are presented in this section. The first model is the Discrete-State Continuous-Time Markov Process adopted from [22] and the second model is the 2-state Markov Modulated Poisson Process provided in [23], which is used to approximate the discrete-state continuous-time Markov process. The videophone sources provided by Maglaris *et al.* [22] is assumed to be generating images at 30 frames/sec. with each frame consists of approximately 250000 digitally coded pixels. Conditional replenishment compression algorithm is used which encodes and transmits the difference between pixel levels of subsequent frames when the difference exceeds a given threshold. By measuring the output rate of the coder, they have obtained the mean, the variance and the autocovariance function of the video coder output rate. A discrete-state continuous-time Markov process describing the coder output rate is derived by matching the three statistical characteristics. It should be noted that the model that they have derived is applicable to video scenes with relatively uniform activity level only. No abrupt change of scene or error recovery procedure is involved in their image testing sequence which may cause the coding rate to increase suddenly. For convenience, the unit for the coder output rate is converted from bit/pixel (which Maglaris *et al.* used) to packets/second (which is convenient for matrix-analytic algorithm).

5.2.1 Model A : Discrete-State Continuous-Time Markov Process

In this model that Maglaris *et al.* have provided, the output rate from the coder

is assumed to be quantized into finite discrete levels and the coder generates packets of fixed length at pre-defined fixed (deterministic) rate specified at each level. The transitions between the levels are assumed to occur with exponential transition rates which are dependent on the current level. The state space of this discrete finite-state continuous-time Markov process (DS-CTMP) is the set of the quantized levels up to a maximum level. Assuming the output coding rate is given in uniform quantization step of A packets/sec and $M+1$ possible levels $(0, A, \dots, MA)$, the packet generation rate of an individual videophone source has a binomial distribution with mean $E(\lambda)$, variance $C(0)$ and exponential autocovariance $C(\tau)$. The state transition diagram is shown in Fig.5.3. For the binomial model, the following set of equations (which is provided in [22]) should be satisfied :

$$\begin{aligned} \text{Prob.}\{\text{Packet Rate} = mA\} &= \binom{M}{m} p^m (1-p)^{M-m}, & p &= \frac{\beta}{\alpha + \beta} \\ E(\lambda) &= MAp \\ C(0) &= MA^2 p(1-p) \\ C(\tau) &= C(0)e^{-(\alpha+\beta)\tau} \end{aligned} \tag{5.1}$$

The measured mean, variance and autocovariance for the output coder rate provided in [22] is given in bit/pixel. For the coder which generates 30 frames/sec. and 250000 pixel/frame, 1 bit/pixel is equivalent to 7.5 Mbit/sec.. Assuming the video packet size is k Mbits, then for an individual videophone source,

$$\begin{aligned} E(\lambda) &= 3.9/k \text{ packets/sec.} \\ C(0) &= 3.015/k^2 \text{ (packet/sec.)}^2 \\ \alpha + \beta &= 3.9 \text{ /sec} \end{aligned} \tag{5.2}$$

The parameters A , α , and β can be evaluated given the above equations and constants with an arbitrary number of quantization levels M for individual videophone source. The number of quantization levels should be chosen large enough to obtain a good

representation of the actual output rate of the coder. As $M \rightarrow \infty$, the process resembles the output coder rate which is a continuous-state continuous-time process. A minimum of $M = 20$ is suggested in [22]. It should be noted that the number of states grow enormously as the number of aggregated videophone sources increases.

Figure 5.3 State transition diagram for videophone source — Model A

5.2.2 Model E : Markov Modulated Poisson Process

A Markov Modulated Poisson Process (MMPP) is a doubly stochastic Poisson process where the rate process is determined by the state of a continuous-time Markov chain. It is a correlated non-renewal process and it is a special case of the Versatile Markovian Point Process (VMPP) treated in [24]. This process is chosen as the approximating process due to its analytic simplicity and the desire for versatility. It is chosen in [23] to approximate the aggregated voice arrival process where each component voice process is approximated as a renewal process with deterministically spaced packets during talk spurts followed by an exponentially distributed silence period. The aggregated arrival process is a complex non-renewal process.

Similarly, the binomial model presented previously (the discrete-state continuous-time Markov process) can be treated as M identical independent *minisources*, each alternating between transmitting nothing (OFF state) and A packet/sec. (ON state) according to a Bernoulli distribution. As shown in Fig.5.4, a minisource turns ON with exponential rate β and OFF with rate α , which is similar to a single voice source. The aggregate rate out of M minisources corresponds exactly to quantizing the aggregate bit

2

MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS
STANDARD REFERENCE MATERIAL 1010a
(ANSI and ISO TEST CHART No. 2)

rate of single videophone source into M levels. For L independent videophone sources, this corresponds to ML independent identical minisources with each of the minisources is considered as renewal process.

Figure 5.4 Minisource model — Videophone source

In similar respect, a 2-state MMPP (see Fig.5.5) is chosen for the approximation of the aggregate bit rate of L independent videophone sources with each videophone source is composed of M minisources (see Fig.5.6). The underlying Markov chain of the 2-state MMPP has mean sojourn times in state 1 and state 2 as r_1^{-1} and r_2^{-1} respectively. The arrival process is Poisson with rate λ_j when the chain is in state j ($j = 1, 2$).

Figure 5.5 The 2-State MMPP model

For the 2-state MMPP to approximate the superposition of minisources, four parameters of the 2-state MMPP, namely, r_1 , r_2 , λ_1 and λ_2 are evaluated by matching several statistical characteristics with the superposition of minisources, which include :

Figure 5.6 Approximation of aggregate videophone sources with MMPP

1. the mean packet arrival rate in $(0, t]$.
2. the variance-to-mean ratio (Index of Dispersion) of the number of packet arrivals in $(0, t_1)$
3. the long term variance-to-mean ratio of the number of packet arrivals, and
4. the third central moment of the number of arrivals in $(0, t_2)$

Expressions for the above minisources' statistical characteristics are obtained from renewal theory results since each minisource is considered as a renewal process. These expressions and those of the 2-state MMPP, and the procedures for obtaining the four parameters of the 2-state MMPP are given in Appendix G. For detailed derivations, one should refer to [23]. It should be noted that in the matching of the third central moment and the variance-to-mean ratio, there is freedom in choosing different time points, t_1 and t_2 . They should be chosen such that the overall variance-to-mean vs. time curve of the 2-state MMPP is a good fit to that of the superposition of minisources.

For the 2-state MMPP, the equilibrium probability vector $\underline{\pi}$, which denotes the steady state probability distribution of the state of the arrival process, is given by

$$\underline{\pi} = \frac{1}{r_1 + r_2} (r_2, r_1) \quad (5.3)$$

and define

$$Q^* = \begin{bmatrix} -r_1 & r_1 \\ r_2 & -r_2 \end{bmatrix} \quad (5.4)$$

and

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \quad (5.5)$$

where Q^* is the finite, irreducible generator of the 2-state MMPP and Λ is the diagonal matrix of Poisson arrival rates. These matrices and vector characterize the 2-state MMPP and they are to be used in the queueing analysis described in the following section. The advantages of the characterization of the superposition of videophone minisources as MMPP are that it can be fed into any system once the parameters of the process are obtained and the dimensionality of the MMPP is kept small which makes analytic solution more feasible.

5.3 Queueing Performance of Statistical Multiplexer

Often in ISDN, the statistical multiplexer has to exercise access control to guarantee the grade of service. In this section, the performance of a statistical multiplexer with inputs consisting of high priority (videophone) streams and low priority (data) streams are discussed. The streams are multiplexed onto c parallel synchronized high-speed transmission channel with non-preemptive priority strategy applied at the packet transmission level. The following assumptions are made for the queueing model (see Fig.5.7) :

1. The high priority videophone sources and the low priority data sources are all divided into fixed length packets.
2. Channel time is divided into contiguous slots of a packet transmission time. Packet transmissions are synchronized to the slots and all c channels are synchronized and identical. All c channels have the same destination.
3. The packets arrive just prior to a slot. The high priority packets arrive according to a 2-state Markov Modulated Poisson Process derived in the previous section (Sect.

Figure 5.7 Queueing model for statistical multiplexer at packet transmission level

5.2.2). The low priority packets's arrival is independent and identically distributed with

$$\alpha_k = \Pr [Y_n = k], \quad k \geq 0, n \geq 1 \quad (5.6)$$

where Y_n is the number of packets arrived at the beginning of the n^{th} slot. In the analysis, the low priority packet arrival process is chosen to be a Poisson process.

4. The system is being observed just prior to the transmission of a slot.
5. The event of high priority packet arrival is independent of those of the low priority packet arrival.
6. The packets join the infinite FIFO queues according to their own priority type when they arrive at the system.

5.3.1 Analysis — MMPP/D/c Queueing System for High Priority Packets

With synchronized transmission and priority strategy, the high priority packet queueing performance would not be affected by the low priority packets and hence it can be

analyzed separately as a MMPP/D/c queueing system. The arrival process employed in the study is a 2-state MMPP with infinitesimal generator Q^* and diagonal matrix of Poisson arrival rates Λ . The analysis shown here is applicable for an m -state MMPP.

The system being observed at the beginning of the n^{th} slot prior to transmission can be described by the state vector

$$\underline{H}S_n = ({}_H Q_n, H_n), \quad (5.7)$$

where

$$\begin{aligned} {}_H Q_n &= \text{the number of high priority packets waiting in the high priority queue} \\ H_n &= i, 1 \leq i \leq m : \text{arrival process is in the } i^{\text{th}} \text{ state of MMPP} \end{aligned} \quad (5.8)$$

The sequence $\{\underline{H}S_n, n \geq 1\}$ is a 2-dimensional, first-order Markov chain.

Denotes $N(t)$ and $J(t)$, $t \geq 0$ as the number of packet arrivals in $(0, t]$ and the state of the Markov process Q^* at time t respectively. Then $\{N(t) = k, J(t) = i, t \geq 0, k \geq 0, i = \{1 \dots m\}\}$ is a Markov process. Define the joint probability of the number of packet arrival and the arrival state in $(0, t]$ as

$$P_{i,j}(v, t) = \Pr[N(t) = v, J(t) = j \mid N(0) = 0, J(0) = i], \quad 1 \leq i, j \leq m. \quad (5.9)$$

and the duration of one packet transmission as T . Then the matrix $P(v, T) = \{P_{i,j}(v, T)\}$ for $v \geq 0$ can be obtained by numerical integration of the following set of differential equations :

$$\begin{aligned} P'(0, T) &= P(0, T)[Q^* - \Lambda] \\ P'(v, T) &= P(v, T)[Q^* - \Lambda] + P(v-1, T)\Lambda \quad \text{for } v \geq 1 \end{aligned} \quad (5.10)$$

with initial conditions $P(v, 0) = \delta_{v0}I$, for $v \geq 0$ [25].

Define the joint steady state probability for the high priority packets

$$x_{k,l} = \Pr[{}_H Q_n = k, H_n = l], \quad k \geq 0, 1 \leq l \leq m \quad (5.11)$$

and the steady probability vectors as

$$\underline{x}_k = [x_{k,1}, x_{k,2}, x_{k,3}, \dots, x_{k,m}], \quad k \geq 0 \quad (5.12)$$

and

$$\underline{x} = [\underline{x}_0, \underline{x}_1, \underline{x}_2, \dots], \quad (5.13)$$

then,

$$\underline{x} \mathbf{P} = \underline{x}, \quad \underline{x} \underline{e} = 1, \quad (5.14)$$

where

$$\underline{e} = (1, 1, 1, 1, \dots)^T, \quad (5.15)$$

and the stochastic transition probability matrix $\mathbf{P}$ is given by

$$\mathbf{P} = \begin{matrix} & \begin{matrix} \langle 0 \rangle \\ \langle 1 \rangle \\ \vdots \\ \langle c-1 \rangle \\ \langle c \rangle \\ \langle c+1 \rangle \\ \langle c+2 \rangle \\ \langle c+3 \rangle \\ \vdots \end{matrix} & \begin{bmatrix} P(0,T) & P(1,T) & P(2,T) & P(3,T) & \dots \\ P(0,T) & P(1,T) & P(2,T) & P(3,T) & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ P(0,T) & P(1,T) & P(2,T) & P(3,T) & \dots \\ P(0,T) & P(1,T) & P(2,T) & P(3,T) & \dots \\ 0 & P(0,T) & P(1,T) & P(2,T) & \dots \\ 0 & 0 & P(0,T) & P(1,T) & \dots \\ 0 & 0 & 0 & P(0,T) & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \end{matrix} \quad (5.16)$$

The first $(c+1)$ rows of $m \times m$ blocks, i.e. $P(0,T)$, $P(1,T)$, $P(2,T)$, ... in $\mathbf{P}$, are identical and the rest is a staircase pattern obtained by shifting the rows one block to the right starts from the $(c+1)^{\text{st}}$ row of blocks. The matrix $\mathbf{P}$ is then further partitioned into square blocks of order $c \times m$ in the natural manner such that numerical solutions for

such matrix_type is known. The further partitioned matrix is given by

$$\mathbf{P} = \begin{matrix} \underline{0} \\ \underline{1} \\ \underline{2} \\ \underline{3} \\ \vdots \end{matrix} \begin{bmatrix} B_0 & B_1 & B_2 & B_3 & \cdots \\ A_0 & A_1 & A_2 & A_3 & \cdots \\ 0 & A_0 & A_1 & A_2 & \cdots \\ 0 & 0 & A_0 & A_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \quad (5.17)$$

where B_i 's, A_i 's, $i \geq 0$ are of order $c \times m$ and they have the form given by

$$\begin{aligned} B_i &= \begin{bmatrix} P(ic, T) & P(ic+1, T) & \cdots & P((i+1)c-1, T) \\ P(ic, T) & P(ic+1, T) & \cdots & P((i+1)c-1, T) \\ \vdots & \vdots & & \vdots \\ P(ic, T) & P(ic+1, T) & \cdots & P((i+1)c-1, T) \end{bmatrix}, \quad i \geq 0 \\ A_0 &= \begin{bmatrix} P(0, T) & P(1, T) & \cdots & P(c-1, T) \\ 0 & P(0, T) & \cdots & P(c-2, T) \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & P((i+1)c-1, T) \end{bmatrix}, \\ A_i &= \begin{bmatrix} P(ic, T) & P(ic+1, T) & \cdots & P((i+1)c-1, T) \\ P(ic-1, T) & P(ic, T) & \cdots & P((i+1)c-2, T) \\ \vdots & \vdots & & \vdots \\ P((i-1)c+1, T) & P((i-1)c+2, T) & \cdots & P(ic, T) \end{bmatrix}, \quad i \geq 1 \end{aligned} \quad (5.18)$$

The new partitioned transition probability matrix $\mathbf{P}$ shows that the Markov chain at hand is of "M/G/1 Type" and $\mathbf{P}$ is an irreducible stochastic matrix. Here the set of states $\underline{l} = \{(kc + k', j), 0 \leq k' < c, 1 \leq j \leq m\}$ is called the level $\underline{k}$. In Appendix H, the algorithm for solving the invariant probability vector $\underline{x}$ is given. Details of the algorithm should be referred to [24] and [25].

5.3.2 Analysis — Approximation of Queueing Performance for Low Priority Packets

An exact analysis of the low priority packet's queueing performance requires the knowledge of the number of idle servers left to serve the low priority packets from one time slot to another. For the system having deterministic service time, the knowledge of idle servers' availability is disclosed in the high priority queue length information. For example, if the high priority queue has i packets waiting prior to the transmission of the n^{th} slot, then the number of servers (transmission channels of the multiplexer) available to the low priority packets at n^{th} slot is $\max(c-i, 0)$. These information is in fact readily provided in the transition probability matrix $\mathbf{P}$. Exact analysis is difficult due to the infinite dimension of $\mathbf{P}$. An approximation is derived by restricting the knowledge about the number of high priority packet waiting in the queue to a certain limit for the low priority packets. In the analysis, this upper limit on the knowledge of the number of high priority packet waiting is denoted as $\bar{q}$, $\bar{q} \geq c$; the information for high priority queue length that is beyond $\bar{q}$ is aggregated at the upper-bound $\bar{q}$.

A. Approximation 1

The low priority packet queueing system observed just prior to the transmission of the n^{th} slot is described *approximately* by the state vector

$$\underline{L}S_n = ({}_LQ_n, {}_H\tilde{Q}_n, H_n), \quad (5.19)$$

where

${}_LQ_n$ = the number of low priority packets waiting in the low priority queue

$${}_H\tilde{Q}_n = \begin{cases} i, & 0 \leq {}_HQ_n < \bar{q} \\ \bar{q}, & {}_HQ_n \geq \bar{q} \end{cases}$$

$H_n = i, 1 \leq i \leq m$: high priority packet arrival process is in i^{th} state of MMPP

${}_HQ_n$ = the number of high priority packets waiting in the high priority queue

(5.20)

The sequence $\{\underline{L}S_n, n \geq 1\}$ is a 3-dimensional, first-order Markov chain.

The transition probability matrix of the "servers' availability" is in fact a partition of P with the transition probabilities at $\bar{q}$ and beyond is aggregated at the boundary, i.e. at $\bar{q}$. The availability matrix is given as

$$V = \begin{bmatrix} \hat{P}_{0,0} & \hat{P}_{0,1} & \hat{P}_{0,2} & \cdots & \hat{P}_{0,\bar{q}-c} & \hat{P}_{0,\bar{q}-c+1} & \cdots & \hat{P}_{0,\bar{q}-1} & \hat{P}_{0,\bar{q}} \\ \hat{P}_{1,0} & \hat{P}_{1,1} & \hat{P}_{1,2} & \cdots & \hat{P}_{1,\bar{q}-c} & \hat{P}_{1,\bar{q}-c+1} & \cdots & \hat{P}_{1,\bar{q}-1} & \hat{P}_{1,\bar{q}} \\ \vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots \\ \hat{P}_{c-1,0} & \hat{P}_{c-1,1} & \hat{P}_{c-1,2} & \cdots & \hat{P}_{c-1,\bar{q}-c} & \hat{P}_{c-1,\bar{q}-c+1} & \cdots & \hat{P}_{c-1,\bar{q}-1} & \hat{P}_{c-1,\bar{q}} \\ \hat{P}_{c,0} & \hat{P}_{c,1} & \hat{P}_{c,2} & \cdots & \hat{P}_{c,\bar{q}-c} & \hat{P}_{c,\bar{q}-c+1} & \cdots & \hat{P}_{c,\bar{q}-1} & \hat{P}_{c,\bar{q}} \\ 0 & \hat{P}_{c+1,0} & \hat{P}_{c+1,1} & \cdots & \hat{P}_{c+1,\bar{q}-c-1} & \hat{P}_{c+1,\bar{q}-c} & \cdots & \hat{P}_{c+1,\bar{q}-2} & \hat{P}_{c+1,\bar{q}-1} \\ \vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \hat{P}_{\bar{q}-1,1} & \hat{P}_{\bar{q}-1,2} & \cdots & \hat{P}_{\bar{q}-1,c} & \hat{P}_{\bar{q}-1,c+1} \\ 0 & 0 & 0 & \cdots & \hat{P}_{\bar{q},0} & \hat{P}_{\bar{q},1} & \cdots & \hat{P}_{\bar{q},c-1} & \hat{P}_{\bar{q},c} \end{bmatrix} \quad (5.21)$$

where

$$\hat{P}_{i,j} = \begin{cases} \sum_{l=\bar{q}}^{\infty} P(l, T) & \text{if } 0 \leq i \leq c, j = \bar{q} \\ \sum_{l=j}^{\infty} P(l, T) & \text{if } c < i < \bar{q}, j = c + \bar{q} - i \\ \left(\sum_{l=\bar{q}}^{\infty} x_l e \right)^{-1} \sum_{l=\bar{q}}^{\bar{q}+j} x_l e P(\bar{q} + j - l, T) & \text{if } i = \bar{q}, 0 \leq j < c \\ \left(\sum_{l=\bar{q}}^{\infty} x_l e \right)^{-1} \sum_{r=\bar{q}}^{\infty} x_r e \sum_{s=\max(\bar{q}+c-r, 0)}^{\infty} P(s, T) & \text{if } i = \bar{q}, j = c \\ P(j, T) & \text{if } (0 \leq i \leq c, j < \bar{q}) \text{ or} \\ & (c < i < \bar{q}, j < c + \bar{q} - i) \end{cases} \quad (5.22)$$

The matrix $\hat{P}_{i,j}$ is defined as $\hat{P}_{i,j} = \{\hat{P}_{i,j}(r, s), 1 \leq r, s \leq m\}$, $0 \leq i, j \leq \bar{q}$ where

$$\hat{P}_{i,j}(r, s) = \Pr \left[{}_H\tilde{Q}_n = j, H_n = s \mid {}_H\tilde{Q}_{n-1} = i, H_{n-1} = r \right], \quad n \geq 1. \quad (5.23)$$

The followings are defined :

1. The joint steady state probability for low priority packet is defined as

$$y_{i,j,k} = \Pr \left[{}_L Q_n = i, {}_H \tilde{Q}_n = j, H_n = k \right], i \geq 0, 0 \leq j \leq \tilde{q}, 1 \leq k \leq m \quad (5.24)$$

2. The invariant probability vector for the queue length distribution of low priority packets is defined as

$$\underline{y} = (\underline{\tilde{y}}_0, \underline{\tilde{y}}_1, \underline{\tilde{y}}_2, \dots), \quad (5.25)$$

where the partitions are given as

$$\begin{aligned} \underline{\tilde{y}}_k &= (\underline{\hat{y}}_{kc}, \underline{\hat{y}}_{kc+1}, \dots, \underline{\hat{y}}_{(k+1)c-1}), \quad k \geq 0, \\ \underline{\hat{y}}_{kc+i} &= (\underline{y}_{kc+i,0}, \underline{y}_{kc+i,1}, \dots, \underline{y}_{kc+i,\tilde{q}}), \quad 0 \leq i \leq c-1, \text{ and} \\ \underline{y}_{kc+i,j} &= (y_{kc+i,j,1}, y_{kc+i,j,2}, \dots, y_{kc+i,j,m}), \quad 0 \leq j \leq \tilde{q}. \end{aligned} \quad (5.26)$$

The stochastic transition probability matrix for the low priority packet queueing system, $\tilde{\mathbf{P}}$, has similar structure as $\mathbf{P}$ given in Eqn.(5.16), it is then further partitioned and the result is given as

$$\tilde{\mathbf{P}} = \begin{bmatrix} \tilde{B}_0 & \tilde{B}_1 & \tilde{B}_2 & \tilde{B}_3 & \cdots \\ \tilde{A}_0 & \tilde{A}_1 & \tilde{A}_2 & \tilde{A}_3 & \cdots \\ 0 & \tilde{A}_0 & \tilde{A}_1 & \tilde{A}_2 & \cdots \\ 0 & 0 & \tilde{A}_0 & \tilde{A}_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \quad (5.27)$$

which satisfies

$$\underline{y} \tilde{\mathbf{P}} = \underline{y}, \quad \underline{y} \underline{e} = 1 \quad (5.28)$$

with the submatrices $\tilde{B}_i$'s, $\tilde{A}_i$'s, $i \geq 0$ are of order $c \times m \times (\bar{q} + 1)$ and they have the forms given by

$$\begin{aligned} \tilde{B}_i &= \begin{bmatrix} C_{0,ic} & C_{0,ic+1} & \cdots & C_{0,(i+1)c-1} \\ C_{1,ic} & C_{1,ic+1} & \cdots & C_{1,(i+1)c-1} \\ \vdots & \vdots & \ddots & \vdots \\ C_{c-1,ic} & C_{c-1,ic+1} & \cdots & C_{c-1,(i+1)c-1} \end{bmatrix}, \quad i \geq 0 \\ \tilde{A}_i &= \begin{bmatrix} C_{c,ic} & C_{c,ic+1} & \cdots & C_{c,(i+1)c-1} \\ C_{c,ic-1} & C_{c,ic} & \cdots & C_{c,(i+1)c-2} \\ \vdots & \vdots & \ddots & \vdots \\ C_{c,(i-1)c+1} & C_{c,(i-1)c+2} & \cdots & C_{c,ic} \end{bmatrix}, \quad i \geq 0 \end{aligned} \quad (5.29)$$

and

$$C_{r,s} = \Delta(\underline{k}) \cdot \mathbf{V}, \quad 0 \leq r \leq c, \quad s \geq 0 \quad (5.30)$$

where $\Delta(\underline{k})$ is a diagonal matrix of order $(\bar{q} + 1) \times m$ with the diagonal elements form a vector $\underline{k}$ given by

$$\underline{k} = [\underline{k}_0, \underline{k}_1, \dots, \underline{k}_{\bar{q}}] \quad (5.31)$$

where

$$\underline{k}_{v,(0 \leq v \leq \bar{q})} = \begin{cases} \alpha_w \cdot \underline{e}^T & \text{if } w \geq 0 \\ \underline{0} & \text{otherwise} \end{cases} \quad (5.32)$$

with

$$\begin{aligned} \underline{e}^T &= \text{row vector of 1's of size } m, \\ \underline{0} &= \text{row vector of 0's of size } m, \\ w &= r - \min(v, c) + \min(c - r, \min(v, c)), \quad \text{and} \\ \alpha_w &= \text{Prob. of } w \text{ low priority packets arrive in time } T. \end{aligned} \quad (5.33)$$

Each $C_{r,s}$, $0 \leq r \leq c$, $s \geq 0$ matrix describes the possible transitions for the low priority packets under the constraint of servers' availability and the possible number

of packets (low priority) arrival. Since $\tilde{\mathbf{P}}$ indicates that the Markov chain is again of M/G/1 Type, the steady state probability distribution can be readily obtained following the procedure outlined in Appendix A.

B. Approximation 2

In Approximation 1, the applicability of the solution is severely limited due to the size of the matrices $\tilde{B}_k$'s and $\tilde{A}_k$'s, $k \geq 0$ etc. are of order $c \times m \times (\bar{q}+1)$. Approximation 2 is a further simplification of Approximation 1 by retaining only the partial information of the high priority queue length. The information about the arrival states of the high priority packet is aggregated. It is hoped that this simplification, which reduces the matrices' sizes to $c \times (\bar{q}+1)$, will still provide a reasonable approximation for the low priority packet's queuing performance. Since the derivation for the various matrices are similar to those given in Approximation 1, readers should refer to Appendix I for details.

5.4 Numerical Examples and Discussion

5.4.1 MMPP vs. DS-CTMP

As mentioned in section 5.2, the videophone source model is adopted from [22]. The measurements that Maglaris *et al.* has obtained shows that a single videophone source output rate has mean of 3.9 Mbps, variance of 3.015 Mbps², and a exponential time constant of 1/3.9 sec for the exponential approximation of its autocovariance function. An arbitrary M -level Discrete State-Continuous Time Markov Process is used by Maglaris *et al.* as the videophone source model, and the DS-CTMP is further decomposed into M minisources. With the aforementioned parameters and an arbitrary number M , the transition rate between the ON-state and the OFF-state (β and α) as well as the packet arrival rate (A) for each M minisources can be obtained. The aggregation of the

minisources is then further approximated by a 2-state MMPP. The results being reported in [22] is obtained with the fluid flow approximation using aggregation of minisources (DS-CTMP) as traffic source. The matrix-analytic analysis employed in the studies is discrete in nature where the queue is observed at every packet transmission time, hence packet size is chosen to be 10k bits per individual videophone source in order to provide a comparable result. The selection of 10k bits per packet is arbitrary and it seems that the results obtained are in general comparable with those reported in [22]. Other packet sizes (which in other words means different durations for one packet transmission) have also been studied, and it is found that matching of statistical characteristics is impossible in this particular settings (i.e. mean rate / videophone source = 3.9 Mbps) when packet transmission time is approximately 15 msec. and beyond.

As described in section 5.2.2, there is freedom in choosing different time points (t_1 and t_2) for matching the statistical characteristics (variance-to-mean ratio (or Index of Dispersion) and third central moment for the number of videophone packet arrival in $(0, t]$) of the 2-state MMPP with the DS-CTMP. The effect of choosing different t_1 and t_2 is not well known. Fig.5.8 shows the Index of Dispersion curve for the number of videophone packet arrival from a single videophone source. M is chosen as 120, utilization = 0.8 and packet size is 10k bits. These parameters are chosen to match those as reported in [22]. It is shown in Fig.5.8 that the choice of t_1 and t_2 is rather insensitive to the Index of Dispersion, and this agrees with the result reported in [23]. This indicates that the second order statistics (variance) can be matched well between the MMPP and DS-CTMP models without great discrepancy, and also it is rather insensitive to the choice of t_1 and t_2 . With the same set of parameters, the queue length distribution (reported as Prob. of [Queue Length given in msec. > k] where k is buffer size in msec.) given different choices of t_1 and t_2 (unit is msec.) is plotted in Fig.5.9. (The buffer size has units of msec., this can be converted back to number of packets since the

utilization and the mean rate of the videophone source are known.) The results show great discrepancies among different choices of t_1 and t_2 . It is conjectured that the good matching of first and second order statistics may not yet be adequate to result in a good approximate model for correlated sources. Although the third central moment for the number of packet arrival in time t is matched at time point t_2 , it is found during the analysis that the ratio of third central moment-to-mean in time t does not match very well between the MMPP and DS-CTMP. The mismatch is mainly due to numerical problem. However, it is still unknown whether the close match of third central moment in time t will yield good approximation and this is left for further studies. It should be noted that in Fig.5.9, it shows the trend of how the choice of t_1 and t_2 affect the queue length distribution. There are mainly four groups ($[a]$, $[b\ c]$, $[d\ e]$ and $[f\ g\ h]$ where $a, b, \dots, h$ are curve labels) shown in Fig.5.9, and each member in a group has the same t_2 . The higher the value of t_2 , the higher the queue length distribution, and within each group, the smaller the value t_1 , the higher the queue length distribution among the same group.

The choice of t_1 and t_2 seems to be quite arbitrary, $t_1 = 500$ msec. and $t_2 = 4000$ msec. are chosen for further analysis since they show a close match for the queue length distribution of single videophone source in the analysis when it is compared to the result reported in [22]. Fig.5.10 shows the average queue size (given in terms of msec.) for different utilizations with different number of videophone sources aggregated in the single channel multiplexer (i.e. $c = 1$). For each videophone source generates at a mean rate of 3.9 Mbps, the aggregation of multiple videophone sources would increase the mean aggregated rate, hence the packet size is chosen to be multiple of 10k bits such that the duration for one packet transmission is kept the same. Each videophone source is composed of 1000 ($= M$) minisources instead of 120. The choice of $M = 1000$ is arbitrary. The result for DS-CTMP shown in [22] is repeated here in Fig.5.10 for comparison. The average queue length obtained from the MMPP is shown to be very

**Figure 5.8 Index Of Dispersion curve of single videophone source
 with stat. matching at different time points**

Figure 5.9 Queue length distribution of single videophone source with stat. matching at different time points (No. of servers in system [c] = 1)

Figure 5.10 Comparison of average queue size for videophone packets under different utilization (No. of servers $[c] = 1$)

Figure 5.11 Comparison of queue length distribution for videophone packets (No. of servers [c] = 1)

close as compared to that of DS-CTMP under various utilizations. It also shows that the more videophone sources is aggregated, the smaller the average queue size and hence better performance. With the same set of parameters and at a utilization of 0.8, the queue length distribution is shown in Fig.5.11. Although the queue length distribution of a single videophone source matches quite closely between the MMPP and DS-CTMP models, the others show great discrepancies.

5.4.2 Approximation of Low Priority (Data) Packet Queueing Performance

The high priority packets, which are to be multiplexed and transmitted in the multiplexer, are the videophone packets generated from the aggregation of 10 videophone sources (1000 minisources per videosource). The multiplexer used in the numerical example here has 3 synchronous channels ($c = 3$). The parameters used for the 2-state MMPP approximating the aggregation of videophone sources are similar to those used in Fig.5.10 and Fig.5.11. The low priority packet have arrival process according to a Poisson process. Each simulation curve reported in the figures (Fig.5.12 to Fig.5.14) is the result of at least 1,000,000 slots (packet transmission) of simulation. Approximations for the low priority packet queueing performance are shown and compared with simulation results.

In Fig.5.12, the high priority packets arrival contributes to 50% of the total system utilization. Two sets of low priority (data) packet arrivals are considered : one set contributes 30% to system utilization which make total system utilization amount to 80%, another set contributes 40% to system utilization which make total system utilization amount to 90%. Comparisons between the approximation methods (Approximation 1 and Approximation 2) under different utilizations are shown. The high priority queue length information that is made known to the low priority packet ($\bar{q}$) is chosen to be 4. As shown in Fig.5.12, Approximation 1, which includes the details of the arrival states of

high priority packets (2-state MMPP), is a very good approximation under both sets of arrivals (i.e. util. = 0.3 and 0.4). Approximation 2, which aggregates all the information about high priority packet arrival states, is very poor. It indicates that if the high priority packet arrival process is MMPP, then the arrival state information is crucial and cannot be aggregated. This will severely limit the application of the approximation for low priority packet queueing performance since the dimensionality of the model, which is mentioned in section 5.3.2.2, cannot be reduced without the sacrifice of the accuracy of the approximation. It also indicates that a good and mathematically tractable approximation method is very important in accurately approximating the queueing performance for the low priority packet, which is subjected to the influence of high priority *correlated* packet arrival. The solid line in Fig.5.12 (as well as in Fig.5.13 and Fig.5.14) represents the queue length distribution of the high priority packets generated from an aggregation of 10 videophone sources. It should be noticed (as shown in Fig.5.12) that the application of access control will cause the low priority data packets to suffer extensive delays due to the building up of a long queue. Positive correlation in arrival process has strong influence and it is expected that if simple Poisson process is used for approximation, it will severely under-estimate the queueing behaviour.

Fig.5.13 and Fig.5.14 show the effects of $\bar{q}$ on the accuracy of approximation under different system utilizations. The approximation method used is "Approximation 1" described in section 5.3.2.1. In Fig.5.13, high priority packet arrival contributes to 50% of the total system utilization. Two sets of low priority packet arrival, one set contributes 30% to the total utilization and the other contributes 40% to the utilization, are analyzed and compared with simulations. In each set of packet arrival, approximations with $\bar{q} = 3$ and $\bar{q} = 4$ are compared. With $\bar{q} = 4$, the approximations are very close to the simulation results. It is indicated in the figure that 95% of the time, the buffer size of high priority packet queue is not greater than approximately 16 msec., which is

Figure 5.12 Comparisons of Approximations for queue length distribution for data packets (Utilization for videophone sources = 0.5, $c = 3$, $\bar{q} = 4$)

Figure 5.13 Effects of $\bar{q}$ (queue length information) on the accuracy of approximation (Utilization for videophone sources = 0.5, $c = 3$)

Figure 5.14 Effects of $\bar{q}$ (queue length information) on the accuracy of approximation (Utilization for videophone sources = 0.65, $c = 3$)

equivalent to approximately 4 packets (1 packets transmission time is 3.85 msec. under 50% utilization). Similar result is shown in Fig.5.14 where good approximations can be obtained with $\bar{q} = 7$ under system utilizations of 95% and 85% (65% of the system utilization is contributed by the high priority packet arrival). It is clearly shown in the figure that the buffer size is not greater than 7 for 95% of the time (i.e. Log_{10} of $\text{Prob}(\text{Queue length} > k) = -1.3$ where k is buffer size given in no. of packets).

5.5 Conclusions

In this chapter, a preliminary study on the suitability of a 2-state Markov Modulated Poisson Process as an approximation model for videophone traffic source is presented. An approximation based on matrix-analytic approach is also derived to approximate the queueing behaviour of low priority packets when this class of traffic is subjected to the influence of high priority traffic in a multi-channel multiplexer. The videophone source model is adopted from Maglaris *et al.* where they use a Discrete State-Continuous Time Markov Process to approximate the videophone source. Since the DS-CTMP consists of many states, a 2-state MMPP is used to approximate the aggregation of videophone sources for its simplicity and versatility in capturing correlation effects.

In the first part of the study, it is shown that the variance and mean for the MMPP can be matched to the DS-CTMP without difficulty. The index of dispersion curve is also insensitive to the different time points, t_1 and t_2 , chosen for the statistical characteristics matching. However, different values of t_1 and t_2 do affect the queue length distribution significantly. In approximating the mean queue length, the 2-state MMPP provides comparable results to those reported by Maglaris *et al.*, but the comparison between queue length distribution is not satisfactory.

Although the studies show the inadequacy of 2-state MMPP as videophone source model, the results obtained is only superficial and since MMPP provides the versatility desired in incorporating correlations in a model and at the same time the simplicity desperately needed for a mathematically tractable model, further investigation should be engaged to explore its potentials. During the course of research, the 2-state MMPP has also been used to approximate the continuous-state first-order autoregressive Markov model reported in [22]. Although the index of dispersion curve can be matched quite well, there is difficulty in matching the overall third central moment. Difficulty in matching also arises when the duration of one packet transmission time is greater than 10 msec. (one frame time in original model is 33 msec.). The result obtained seems to be not convincing and it is suspected that the one of the reasons for the mismatch is due to the nature of periodicity of the continuous-state first-order autoregressive Markov model, which the MMPP seems to have difficulty in matching. Also reported in [26] is a model capable of modelling abrupt changes in output coder rate where the single layer DS-CTMP in [22] is modified to a 2-layer DS-CTMP in [26]. Each layer approximates the coder output rate in high and low activity respectively. Two 2-state MMPPs have also been used to approximate each layer of DS-CTMP respectively. Results obtained are compared to those reported in [26]. Similar phenomenon is observed where the queue length distribution does not match very well and it is also dependent upon the value of t_1 and t_2 .

The queueing performance of a multi-channel multiplexer, which has input consisting of high priority correlated traffic (modelled as MMPP) and low priority traffic (modelled as Poisson Process), is analyzed via matrix-analytic approach in the second part of the study. It is shown that the low priority packets will suffer long delay since they can only be serviced when there is no high priority packets waiting. The correlation in high priority packet arrival also significantly affects the queueing performance.

Approximation 1 is an accurate approximation for the queueing performance of the low priority packet as compared with Approximation 2. This indicates that the information about the arrival states of the high priority packets is crucial and cannot be ignored. Unfortunately, since the arrival state information has to be retained, the applicability of Approximation 1 is very limited due to the vast dimension required for analysis.

Chapter 6

Conclusions

6.1 Conclusions

The three queueing systems being studied in this thesis can be modelled as "Markov process/chain of M/G/1 Type" or special cases of "Markov process/chain of GI/M/1 Type". Matrix analytic methodology is found to be a useful tool in solving these types of queueing system. When the system is properly modelled, the numerical procedures of the matrix analytic solutions are found to be stable, and the solutions, especially probability distributions, can generally be obtained with ease when compared with classical methods.

This methodology uses the special block structures of the transition matrix which reduces dimensionality problem, hence it helps in analyzing some queueing systems where previous methods are known to be difficult to apply. Since the three queueing models presented in this thesis are the basic models of M/G/1 or GI/M/1 Type, the solution techniques are rather standardized.

There are still some difficulties with this matrix analytic method. The main difficulty is the low convergence rate in some of the iteration processes for solving various first passage time distribution matrices or the rate matrix R . The slow convergence is mainly due to the skewness of the parameter values and occur typically at high system utilization (approximately ≥ 0.8). Despite the disadvantages, this matrix analytic

method is in general superior to the classical analytic method in obtaining the probability distributions. The conclusions for the three queueing system studied are given in the following paragraphs.

Selective Repeat-Automatic Repeat Request Scheme

Obtaining an *exact* solution without the consideration of the number of states (corresponding to the history of transit queue length) which grows exponentially with N is impossible. Although with the aid of matrix-analytic approach where the dimensionality constraint (2^N) is relaxed from the previous analysis of 2^4 to 2^7 , $N = 7$ may still not be high enough for the analysis to be useful. The applicability of the exact solution for a satellite system employing SR-ARQ scheme or a high-speed communications network employing edge-to-edge SR-ARQ scheme in which N will be considerably large ($N > 7$) is void. The proposed approximate model which is based on partial history of the system is able to obtain an accurate queue length distribution of SR-ARQ scheme under the condition in which the previously known methods fail to provide reasonable results or are difficult to apply due to numerical problems. The approximate model also allows the ACK time N to be extended to approximately 100 time slots ($N = 100$), which makes the analysis for the system with long ACK time possible. Possible extensions to the approximate model to solve the system with correlated block errors and hybrid of FEC/SR-ARQ scheme are also proposed and analyzed without incurring undue increase in numerical complexity.

Access Control Strategies in Wide-band Integrated Network

Analytic methods for solving two-priority M/M/c queueing system with three different types of priority policy (i.e. three types of access control strategies) are presented. The numerical examples show that the selection between preemptive priority (PP) policy and non-preemptive priority (NPP) policy should be dependent on the set of parameter

values given (arrival rates and service rates etc.) in order to achieve a better overall performance for both Type 1-SR and Type 2-SR. General rules of choosing PP policy or NPP policy given the set of parameter values cannot be set due to the fact that thorough study is difficult to achieve during the course of research. Numerical difficulties are encountered when the set of parameter values is skewed. Although the analysis for system with preemptive-priority policy is easier than for the non-preemptive case, using preemptive-priority policy as an approximation of non-preemptive policy is sometimes inappropriate. The attempt for such an approximation shall be accompanied by a proper choice of parameter values. It is also shown that the state spaces for the two NPP strategies are complex, where the complexity stems from the non-preemptive nature and flexible seizures of multiple servers for Type 1-SR. Numerical solutions are made possible due to special tridiagonal structure of the rate matrix R where individual partition in R can be computed recursively.

Videophone Source Modelling and Related Queueing Performance

The preliminary study on the suitability of a 2-state Markov Modulated Poisson Process (MMPP) as an approximation model for videophone traffic source show that the steady state variance, mean, and the Index of Dispersion curve for the 2-state MMPP can be matched with the DS-CTMP (the original videophone source model) without difficulty. However, the selection of time points t_1 and t_2 in the statistical characteristics matching should be careful since the queue length distribution is rather sensitive to the different selections. It seems that the good matching of first and second order statistics may not yet be adequate to result in a good approximate model for correlated source. Numerical problems discovered in matching the ratio of third-central-moment to the mean in time t still requires further attention. Apparently, the 2-state MMPP used in this study is shown to have failed in approximating the videophone sources. However, due to the MMPP's simplicity and versatility in capturing correlation effects, it may still

be a potential model for the modelling of traffic sources with correlation.

In the approximation of the queueing performance of a multi-channel multiplexer, where the input consists of high priority correlated traffic (2-state MMPP) and low priority traffic (Poisson Process), the information about the arrival states of the high priority packets is crucial. However, retaining of the arrival state information lead to the vast dimension required for analysis and hence the applicability of the approximation method is rather limited. The analysis also indicates that the correlation in high priority packet arrival will significantly affect the queueing performance and thus proper modelling of correlated arrival process is important such that actual system performance is analyzed with accuracy.

6.2 Suggestions For Future Work

Selective Repeat-Automatic Repeat Request Scheme

The SR-ARQ scheme studied is in fact a simple basic SR-ARQ scheme. With the approximate model, more sophisticated ARQ scheme such as the extensions presented can be analyzed. Although additional dimensions for the system state description are required as the ARQ scheme become more sophisticated, and the increase in dimension will severely affect the applicability of the analysis due to limited computing resources, yet it is noticed that the matrix blocks in the transition probability matrix $\mathbf{P}$ is very sparse. With the sparse matrix technique and the approximate model, some other sophisticated ARQ schemes may be analyzed. An example is the application of the SR-ARQ scheme in the ISDN where the ACK time is not fixed at N time slots. This situation arises such as when there are 1) more than one feedback channel available for transmitting the ACK/NACK messages, or 2) fluctuations in delay due to congestion within the network. Another suggestion for further studies is the combination of FEC/SR-ARQ scheme with

correlated block error rate. Although this combination will add two extra dimensions to the state description, it is believed that the sparseness of matrix block can make use of the sparse matrix technique and make the numerical analysis feasible.

Access Control Strategies in Wide-band Integrated Network

In the case of PP policy, the analysis is capable of solving an arbitrary c_2 , $0 \leq c_2 \leq c$. The analysis for the case of NPP policy is only capable of solving for $c_2=0$ or $c_2=c$. It should be noticed that the rate matrix R in these two cases have the special tridiagonal structure which makes the analysis possible. More thorough understanding of the significance and the property of the rate matrix R are required such that the analysis for NPP system may become feasible with $0 < c_2 < c$. Another possibility for solving the queueing system with non-preemptive priority policy and with $0 < c_2 < c$ is to seek and use a different form of state description. It seems that with the state description provided in the thesis, the matrix R will not have regular structure and solving for the infinite size R would be rather difficult.

Videophone Source Modelling and Related Queueing Performance

In order to obtain a better approximation of the videophone source by the 2-state MMPP, further investigations are required. In the preliminary study, there are few unsolved problems on which further investigation is worthwhile. It is not known if the mismatch between the DS-CTMP and the 2-state MMPP is due to (a) the nature of the 2-state MMPP, which simply is not capable of approximating the videophone source; or (b) the incorrect matching of time points t_1 and t_2 , which then requires knowledge about the significance of these time points in the statistical characteristics matching process; or (c) the fact that matching the first few statistical characteristics is simply not adequate, which then requires other forms of statistical characteristics be incorporated in the characteristics matching process. Although the approximation of the 2-state

MMPP to the DS-CTMP (in terms of queue length distribution) is not very convincing, the 2-state MMPP model has provided certain versatility and simplicity in modelling arrival process which has correlation, and at the same time the dimensionality is kept small. In addition, the videophone sources employed in the study is only for rather stable picture scenes. In future Broadband-ISDN, TV/video services are to be provided and they have abrupt fluctuations in the output bit rates. Multi-state MMPP, or other types of VMPP (Versatile Markovian Point Process) which incorporate group arrival process, may still be considered as a good choice for providing accurate approximation for this type of sources while the model is kept mathematically tractable. If the MMPP is an appropriate model, the number of states required to capture abrupt changes in the arrival rates and the method for matching the various statistical characteristics with the actual source requires further investigation.

Appendix A

Matrix Analytic Solution for M/G/1 type Paradigm

The following is the algorithm used to solve the invariant probability vector for a Markov chain of M/G/1 Type with

$$\underline{x} \mathbf{P} = \underline{x}, \quad (\text{A.1})$$

where

$$\underline{x} \underline{e} = 1, \quad \underline{x} = (\underline{x}_0, \underline{x}_1, \underline{x}_2, \dots), \quad (\text{A.2})$$

and

$$\mathbf{P} = \begin{bmatrix} B_0 & B_1 & B_2 & B_3 & . & . \\ C_0 & A_1 & A_2 & A_3 & . & . \\ 0 & A_0 & A_1 & A_2 & . & . \\ 0 & 0 & A_0 & A_1 & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \end{bmatrix} \quad (\text{A.3})$$

B_0 is a $v \times v$ matrix, C_0 is a $m \times v$ matrix, B_k 's ($k \geq 1$) are $v \times m$ matrices and A_k 's ($k \geq 0$) are $m \times m$ matrices and $A = \sum_{k=0}^{\infty} A_k$ is an irreducible stochastic matrix.

1. Find the invariant probability vector $\underline{\pi}$ by iteration where

$$\underline{\pi} A = \underline{\pi}, \quad \underline{\pi} \underline{e} = 1 \quad (\text{A.4})$$

where $\underline{e} = (1, 1, 1, \dots, 1)^T$ is a column vector of an appropriate dimension.

* Power method [27] is suggested in the evaluation of the invariant vector.

2. Find the vector $\underline{\theta}$: $\underline{\theta} = \sum_{k=1}^{\infty} k A_k \underline{e}$.
3. Solve for $\mathbf{G}$ by iteration

$$\mathbf{G} = \sum_{k=0}^{\infty} A_k \mathbf{G}^k \quad (\text{A.5})$$

where $\mathbf{G}$ is a stochastic matrix.

An algorithm, which is based on the Newton-Kantorovich scheme and is proposed by Ramaswami [28], provides faster convergence on the evaluation of the matrix $\mathbf{R}$ ($\mathbf{R} = \sum_{k=0}^{\infty} \mathbf{R}^k A_k$) than ordinary Gauss-Seidel methods (approximately 50% to 70% reduction in CPU time). Slight modifications on the algorithm have been made in order to evaluate the $\mathbf{G}$ matrix (Eqn. (A.5)). Iterative procedures are given as follows :

- i. initialize : $G_0 = \mathbf{0}$ (zero matrix of appropriate size).
- ii. Compute $F(G_k) = G_k - \sum_{v=0}^{\infty} A_v G_k^v$.
- iii. Compute $Z_k = (I - A_1)^{-1}$.
- iv. Compute $Y_k = -F(G_k) + A_1 Z_k + A_2(G_k Z_k + Z_k G_k)$.
- v. Compute $G_{k+1} = G_k + Y_k$.
- vi. Repeat from step (ii) until G_k converges to $\mathbf{G}$.

Note : stopping criterion in the numerical implementation is chosen to be the maximum absolute element-wise difference between the matrix G_k and G_{k+1} is less than 10^{-9} .

4. Check $\pi \underline{\theta} = \rho (< 1)$. (In case of the analysis for SR-ARQ scheme, $\rho = \lambda + \alpha$, ρ is the traffic intensity (or sometimes defined as system utilization).)

5. Find the invariant probability vector $\underline{g}$, where

$$\underline{g} \mathbf{G} = \underline{g}, \quad \underline{g} \underline{e} = 1. \quad (\text{A.6})$$

6. Obtain $\tilde{\mathbf{G}}$: $\tilde{\mathbf{G}} = \underline{e} \underline{g}$.

7. Find vector $\underline{\mu}$:

$$\underline{\mu} = \left(\mathbf{I} - \mathbf{G} + \tilde{\mathbf{G}} \right) \left[\mathbf{I} - \mathbf{A} + \tilde{\mathbf{G}} - \Delta(\underline{\theta})\tilde{\mathbf{G}} \right]^{-1} \underline{e}, \quad (\text{A.7})$$

where $\Delta(\underline{\theta}) = \text{diag}(\theta_0, \theta_1, \theta_2, \dots)$ and $\mathbf{I} = \text{diag}(1, 1, 1, \dots, 1)$ of an appropriate dimension. ($\text{diag}(a, b, c, \dots)$ is diagonal matrix with corresponding diagonal elements $a, b, c, \dots$ etc.)

8. Check if $\underline{g} \underline{\mu} = \frac{1}{1-\rho}$, again $\rho = \lambda + \alpha$ in the analysis for SR-ARQ scheme

9. Compute the following matrices :

$$\begin{aligned} \mathbf{A}_1 &= \sum_{v=1}^{\infty} \mathbf{A}_v \mathbf{G}^{v-1}, \\ \mathbf{A}_2 &= \sum_{v=2}^{\infty} (v-1) \mathbf{A}_v \tilde{\mathbf{G}}, \\ \mathbf{A}_3 &= \sum_{v=1}^{\infty} \mathbf{A}_v = \mathbf{A} - \mathbf{A}_0, \\ \mathbf{B}_1 &= \sum_{v=1}^{\infty} \mathbf{B}_v \mathbf{G}^{v-1}, \\ \mathbf{B}_2 &= \sum_{v=2}^{\infty} (v-1) \mathbf{B}_v \tilde{\mathbf{G}}, \\ \mathbf{B}_3 &= \sum_{v=1}^{\infty} \mathbf{B}_v. \end{aligned} \quad (\text{A.8})$$

10. Compute the following matrices (they should be stochastic) :

$$\begin{aligned} H &= (I - A_1)^{-1} C_0, \\ K &= C_0(I - B_0)^{-1} B_1 + A_1, \\ L &= B_0 + B_1 H. \end{aligned} \tag{A.9}$$

11. Find the invariant probability vectors :

$$\underline{d} = \underline{d} L, \quad \text{and} \quad \underline{k} = \underline{k} K. \tag{A.10}$$

12. Find the following column vectors :

$$\begin{aligned} \underline{h}^* &= (I - A_1)^{-1} \left\{ \underline{e} + (A_3 - A_1 + A_2) (I - G + \tilde{G})^{-1} \underline{\mu} \right\}, \\ \underline{d}^* &= \underline{e} + B_1 \underline{h}^* + (B_3 - B_1 + B_2) (I - G + \tilde{G})^{-1} \underline{\mu}, \\ \underline{k}^* &= \underline{e} + C_0(I - B_0)^{-1} \underline{e} \\ &\quad + \left\{ C_0(I - B_0)^{-1} (B_3 - B_1 + B_2) + (A_3 - A_1 + A_2) \right\} \\ &\quad (I - G + \tilde{G})^{-1} \underline{\mu}. \end{aligned} \tag{A.11}$$

13. Find the probability vectors :

$$\begin{aligned} \underline{x}_0 &= \frac{\underline{d}}{(\underline{d} \underline{d}^*)}, \\ \underline{x}_1 &= \frac{\underline{k}}{(\underline{k} \underline{k}^*)}. \end{aligned} \tag{A.12}$$

14. Check $\underline{x}_0 = \underline{x}_1 C_0 + \underline{x}_0 B_0$.

The above procedures are based on the reference [2]. In order to find the probability vectors, $\underline{x}_i$'s ($i > 1$), the method by Ramaswami [10], which reduces the problem of round-off error propagation and also applicable when A_0 is singular, is employed. It should be noted that in case of the *exact* analysis of SR-ARQ scheme, the matrix A_0 is singular.

15. Find the matrices S_i and $\tilde{S}_i$:

$$S_i = \sum_{v=i+1}^{\infty} A_v G^{v-i-1} (I - A_1)^{-1}, \quad i \geq 1, \quad (A.13)$$

$$\tilde{S}_i = \sum_{v=i}^{\infty} B_v G^{v-i} (I - A_1)^{-1}, \quad i \geq 2,$$

where the matrix S_i is the matrix whose entries are the expected number of visits to the state in level $j + i$ starting from a state in level j before hitting states in level k , where $k < j + i$ ($j > 0$), and $\tilde{S}_i$ is similarly defined except that the starting state is in level 0. Note that in the analysis of SR-ARQ scheme, a state is said to be in level i if $Q_n = i$.

16. Find $\underline{x}_i$'s recursively by

$$\underline{x}_{i+1} = \sum_{j=1}^i \underline{x}_j S_{i+1-j} + \underline{x}_0 \tilde{S}_{i+1}, \quad i \geq 1. \quad (A.14)$$

Note : $\underline{x}_i$'s can be truncated at any level. In the analysis, the truncation level is determined when $\underline{x}_i \underline{e}$ is less than 10^{-8} .

* Note 1 : In Eqn.(A.5), Eqn.(A.8) and Eqn.(A.13), the theoretical upper limit of the summation is infinity. In numerical implementation, this upper limit is determined by

the number of A_i 's. The truncation level of the number of A_i 's in the current analysis is determined when the maximum value among the elements of A_i is less than 10^{-15} .

* *Note 2* : Efficient methods for evaluating Eqn.(A.13), Eqn.(A.14) and error checking are provided in [10]

Appendix B

Elements of transition probability matrix **P** for System with Correlated Block Errors

---

The steady state transmit queue length probability is defined as follows :

$$x_{k,l} = \begin{cases} \Pr[Q_n = k, L_n = \lfloor l/q \rfloor, D_n = l - q \lfloor l/q \rfloor], & k = 0, 0 \leq l \leq (N+1)q - 1 \\ \Pr[Q_n = k, L_n = \lfloor l/q \rfloor + 1, D_n = l - q(\lfloor l/q \rfloor + 1)], & k \geq 1, 0 \leq l \leq Nq - 1 \end{cases} \quad (\text{B.1})$$

where q is the number of possible channel states. ($D_n \in \{0, 1, \dots, q-1\}$ and $q = 2$ is used in subsection 3.5.1.) Then the transition matrix **P** is given by Eqn.(3.11), where B_0 is a $(N+1)q \times (N+1)q$ matrix, C_0 is a $Nq \times (N+1)q$ matrix, B_k 's ($k \geq 1$) are $(N+1)q \times Nq$ matrices and A_k 's ($k \geq 0$) are $Nq \times Nq$ matrices.

The followings are defined first :

1. Q^* = transition probability matrix of channel states.

The (ν, ξ) element of Q^* ($= Q^*(\nu, \xi)$) is the probability that the channel changes from the ν_{th} state to the ξ^{th} state in one slot. Q^* is a stochastic matrix.

2. I = Identity matrix of size q .
3. $\Delta(\alpha) = \text{diag}(\alpha_1, \alpha_2, \dots, \alpha_q)$, where α_v is the block error rate when the channel is in the v^{th} state.
4. $\Delta(\beta) = \text{diag}(\beta_1, \beta_2, \dots, \beta_q)$, where $\beta_v = 1 - \alpha_v$.

The (i, j) element of a submatrix is a $q \times q$ matrix and is then given by :

$$\begin{aligned}
 B_0(i, j) &= \begin{cases} y_1(1 - \frac{1}{N})I, & 0 \leq i \leq N, j = i + 1 \\ y_0(1 - \frac{1}{N})I + (\frac{1}{N})(y_0\Delta(\underline{\alpha}) + y_1\Delta(\underline{\beta}))Q^*, & 0 \leq i \leq N, j = i \\ y_0(\frac{1}{N})\Delta(\underline{\beta})Q^*, & 0 \leq i \leq N, j = i - 1 \\ 0, & \text{otherwise} \end{cases} \\
 &\text{for } 0 \leq i, j \leq N, \\
 B_k(i, j) &= \begin{cases} y_{k+1}(1 - \frac{1}{N})I, & 0 \leq i \leq N, j = i \\ (\frac{1}{N})(y_k\Delta(\underline{\alpha}) + y_{k+1}\Delta(\underline{\beta}))Q^*, & 0 \leq i \leq N, j = i - 1 \\ 0, & \text{otherwise} \end{cases} \\
 &\text{for } k \geq 1, 0 \leq i \leq N \text{ and } 0 \leq j \leq N - 1, \\
 A_0(i, j) &= \begin{cases} y_0(\frac{1}{N-1})\Delta(\underline{\beta})Q^*, & 0 \leq i \leq N - 1, j = i \\ y_0(1 - \frac{1}{N-1})I, & 0 \leq i \leq N - 1, j = i + 1 \\ 0, & \text{otherwise} \end{cases} \\
 &\text{for } 0 \leq i, j \leq N - 1, \\
 A_k(i, j) &= \begin{cases} (\frac{1}{N-1})(y_{k-1}\Delta(\underline{\alpha}) + y_k\Delta(\underline{\beta}))Q^*, & 0 \leq i \leq N - 1, j = i \\ y_k(1 - \frac{1}{N-1})I, & 0 \leq i \leq N - 1, j = i + 1 \\ 0, & \text{otherwise} \end{cases} \\
 &\text{for } k \geq 1 \text{ and } 0 \leq i, j \leq N - 1,
 \end{aligned} \tag{B.2}$$

and $C_0 = [\underline{0} \ A_0]$, where $\underline{0} = (0, 0, \dots, 0)^T$.

Note : The definition of the (i, j) element of the above matrices is the same as indicated in section 3.3.2.

Appendix C

Elements of the transition probability matrix $\mathbf{P}$ for System using FEC/SR-ARQ Scheme

---

The steady state transmit queue length probability is defined as follows :

$$x_{k,l} = \Pr \left[Q_n = k, L_n = \eta, \tilde{L}_{n,R} = \xi \right], \quad (\text{C.1})$$

where $0 \leq \eta \leq N$ for $k = 0$, $1 \leq \eta \leq N$ for $k \geq 1$, $0 \leq \xi \leq \eta$ and

$$l = \begin{cases} \sum_{\nu=0}^{\eta-1} (\nu + 1) + \xi = \frac{\eta(\eta+1)}{2} + \xi, & k = 0, \\ \sum_{\nu=1}^{\eta-1} (\nu + 1) + \xi = \frac{\eta(\eta+1)}{2} + \xi - 1, & k \geq 1. \end{cases} \quad (\text{C.2})$$

Then the transition probability matrix $\mathbf{P}$ is given by Eqn.(3.11), where B_0 is a $m \times m$ matrix, C_0 is a $(m - 1) \times m$ matrix, B_k 's ($k \geq 1$) are $m \times (m - 1)$ matrices and A_k 's ($k \geq 0$) are $(m - 1) \times (m - 1)$ matrices with

$$m = \frac{(N + 1)(N + 2)}{2}. \quad (\text{C.3})$$

The size of the (i,j) elements of submatrices B_0 , B_k 's ($k \geq 1$) and A_k 's ($k \geq 0$) are $(i + 1) \times (j + 1)$, $(i + 1) \times (j + 2)$, and $(j + 2) \times (j + 2)$, respectively. The elements of the

submatrices are given by :

$$\begin{aligned}
 B_0(i, j) &= \begin{cases} y_1(1 - \frac{1}{N})Q_i, & 0 \leq i \leq N, j = i + 1 \\ y_0(1 - \frac{1}{N})I_i + (\frac{1}{N})(y_0T_i + y_1S_i), & 0 \leq i \leq N, j = i \\ y_0(\frac{1}{N})U_i, & 0 \leq i \leq N, j = i - 1 \\ 0, & \text{otherwise} \end{cases} \\
 &\text{for } 0 \leq i, j \leq N, \\
 B_k(i, j) &= \begin{cases} y_{k+1}(1 - \frac{1}{N})Q_i, & 0 \leq i \leq N, j = i \\ (\frac{1}{N})(y_kT_i + y_{k+1}S_i), & 0 \leq i \leq N, j = i - 1 \\ 0, & \text{otherwise} \end{cases} \\
 &\text{for } k \geq 1, 0 \leq i \leq N \text{ and } 0 \leq j \leq N - 1, \\
 A_0(i, j) &= \begin{cases} y_0(\frac{i+1}{N})S_{i+1}, & 0 \leq i \leq N - 1, j = i \\ y_0(1 - \frac{i+1}{N})Q_{i+1}, & 0 \leq i \leq N - 1, j = i + 1 \\ 0, & \text{otherwise} \end{cases} \\
 &\text{for } 0 \leq i, j \leq N - 1, \\
 A_k(i, j) &= \begin{cases} (\frac{i+1}{N})(y_{k-1}T_{i+1} + y_kS_{i+1}), & 0 \leq i \leq N - 1, j = i \\ y_k(1 - \frac{i+1}{N})Q_{i+1}, & 0 \leq i \leq N - 1, j = i + 1 \\ 0, & \text{otherwise} \end{cases} \\
 &\text{for } k \geq 1 \text{ and } 0 \leq i, j \leq N - 1,
 \end{aligned} \tag{C.4}$$

and $C_0 = [\underline{0} \ A_0]$, where $\underline{0} = (0, 0, \dots, 0)^T$.

Note : The definition of the (i, j) element (which itself is a matrix) of the above matrices is the same as indicated in section 3.3.2.

The matrices S_i ($i + 1$ by $i + 1$ matrix), T_i ($i + 1$ by $i + 1$ matrix), U_i ($i + 1$ by i matrix) and Q_i ($i + 1$ by $i + 2$ matrix) are defined as follows :

$$S_i(u, v) = \begin{cases} \beta_F(1 - \frac{u}{i}), & 0 \leq u, v \leq i, v = u \\ \beta_R(\frac{u}{i}), & 0 \leq u, v \leq i, v = u - 1 \\ 0, & \text{otherwise} \end{cases} \tag{C.5}$$

$$\begin{aligned}
T_i(u, v) &= \begin{cases} \alpha_F(1 - \frac{u}{i}), & 0 \leq u, v \leq i, v = u + 1 \\ \alpha_R(\frac{u}{i}), & 0 \leq u, v \leq i, v = u \\ 0, & \text{otherwise} \end{cases} \\
U_i(u, v) &= \begin{cases} \beta_F(1 - \frac{u}{i}), & 0 \leq u \leq i, 0 \leq v \leq i - 1, v = u \\ \beta_R(\frac{u}{i}), & 0 \leq u \leq i, 0 \leq v \leq i - 1, v = u - 1 \\ 0, & \text{otherwise} \end{cases}
\end{aligned} \tag{C.6}$$

and $Q_i = [I_i \ \underline{0}_i]$, where I_i is an identity matrix of size $(i + 1)$, $\underline{0}_i$ is an all zero column vector of size $(i + 1)$, $\beta_F = 1 - \alpha_F$ and $\beta_R = 1 - \alpha_R$.

Note : The index u and v in the above matrices (S_i , T_i , U_i and Q_i) indicates that there are u retransmitted data blocks in the transit queue just after the transmission of the previous slot, and v retransmitted data blocks in the transit queue just after the transmission of the current slot.

Appendix D

Algorithmic Solutions For Non-Preemptive Priority M/M/c Model (Analysis 1 — Sect. 4.3)

Following is a generalization of Miller's methodology [20] to handle multiple flexible seizures of servers for Type 1–SR. The procedures for the invariant probability vector for the M/M/c queueing model where Type 1–SR has absolute non-preemptive priority over Type 2–SR are given below.

The state space of the system is defined in section 4.3 as follows :

1. Let $x_{i,j,k}$ be the state with i Type 1–SR waiting in the Type 1 queue, j Type 2–SR waiting in the Type 2 queue, and k servers serving Type 1–SRs and $(c - k)$ servers serving Type 2–SRs, $i, j \geq 0$, $0 \leq k \leq c$. Then the blocks are defined as

$$H_{i,j} = \{x_{i,j,k} \mid 0 \leq k \leq c\}, \quad i, j \geq 0. \quad (\text{D.1})$$

2. Let $x_{m,n}$ be the state with m server being in service, n Type 1–SR being serviced and no SR is waiting in any of Type 1 or Type 2 queue, $0 \leq m \leq (c - 1)$, $0 \leq n \leq \lfloor \frac{m}{b} \rfloor$, where $\lfloor y \rfloor$ is the largest integer no greater than y . Then the blocks are defined as

$$H_m = \{x_{m,n} \mid 0 \leq n \leq \lfloor \frac{m}{b} \rfloor\}, \quad 0 \leq m \leq (c - 1). \quad (\text{D.2})$$

Therefore, the state space is given as

$$S = \left(\bigcup_{m=0}^{c-1} H_m \right) \cup \left(\bigcup_{i \geq 0} \bigcup_{j \geq 0} H_{i,j} \right). \quad (\text{D.3})$$

The matrices of transition rates between the blocks defined above are denoted as :

$$\begin{aligned} K_i &: H_i \rightarrow H_{i+1}, 0 \leq i \leq c-1 \\ J_i &: H_i \rightarrow H_{i-1}, 1 \leq i \leq c-1 \\ J_c &: H_{0,0} \rightarrow H_{c-1} \\ F_i &: H_i \rightarrow H_{i-b}, b \leq i \leq c-1 \\ F_c &: H_{0,0} \rightarrow H_{c-b} \\ E_i &: H_i \rightarrow H_{i+b}, 0 \leq i \leq c-b-1 \\ E_j &: H_j \rightarrow H_{0,0}, c-b \leq j \leq c-1 \\ M_0 &: H_{i,j} \rightarrow H_{i-1,j}, i \geq 1, j \geq 0 \\ M_i &: H_{0,j} \rightarrow H_{0,j-i}, 1 \leq i \leq b, j \geq 1, j-i \geq 0 \\ X_i &: H_{0,j} \rightarrow H_{c-i+j}, 2 \leq i \leq b, 0 \leq j \leq b-1, c-i+j \geq (c - \text{rem}(c, b)) \\ X'_i &: H_{0,j} \rightarrow H_{c-i+j}, 2 \leq i \leq b, 0 \leq j \leq b-1, c-i+j < (c - \text{rem}(c, b)) \\ L_1 &: H_{i,j} \rightarrow H_{i+1,j}, i, j \geq 0 \\ L_2 &: H_{i,j} \rightarrow H_{i,j+1}, i, j \geq 0 \\ -D_i &: H_i \rightarrow H_i, 0 \leq i \leq c-1 \\ -D_c &: H_{i,j} \rightarrow H_{i,j}, i, j \geq 0 \end{aligned}$$

* $\text{rem}(c, b)$ is the remainder of the division : c/b .

(D.4)

The entry to each of the matrices mentioned above is given below :

$$K_i(p, q) = \begin{cases} \lambda_2 & \text{if } q = p, i \neq c-1, 0 \leq p \leq \lfloor \frac{c}{b} \rfloor, 0 \leq q \leq \lfloor \frac{c+1}{b} \rfloor \\ \lambda_2 & \text{if } q = pb, i = c-1, 0 \leq p \leq \lfloor \frac{c}{b} \rfloor, 0 \leq q \leq c \\ 0 & \text{otherwise} \end{cases}$$

for $0 \leq i \leq c-1$

$$J_i(p, q) = \begin{cases} (i - bp)\mu_2 & \text{if } q = p, i \neq c, 0 \leq p \leq \lfloor \frac{c}{b} \rfloor, 0 \leq q \leq \lfloor \frac{c-1}{b} \rfloor \\ (c - p)\mu_2 & \text{if } q = \frac{p}{b}, i = c, 0 \leq p \leq c, 0 \leq q \leq \lfloor \frac{c-1}{b} \rfloor \\ p\mu_1/b & \text{if } q = \frac{p-1}{b}, i = c, 0 \leq p \leq c, 0 \leq q \leq \lfloor \frac{c-1}{b} \rfloor \\ 0 & \text{otherwise} \end{cases}$$

for $1 \leq i \leq c$

$$F_i(p, q) = \begin{cases} p\mu_1 & \text{if } q = p-1, i \neq c, 0 \leq p \leq \lfloor \frac{c}{b} \rfloor, 0 \leq q \leq \lfloor \frac{c}{b} \rfloor - 1 \\ p\mu_1/b & \text{if } q = \frac{p}{b} - 1, b \neq 1, i = c, 0 \leq p \leq c, 0 \leq q \leq \lfloor \frac{c}{b} \rfloor - 1 \\ 0 & \text{otherwise} \end{cases}$$

for $b \leq i \leq c$

$$E_i(p, q) = \begin{cases} \lambda_1 & \text{if } q = p+1, i < c-b, 0 \leq p \leq \lfloor \frac{c}{b} \rfloor, 0 \leq q \leq \lfloor \frac{c}{b} \rfloor + 1 \\ \lambda_1 & \text{if } q = pb + c - i, i \geq c-b, 0 \leq p \leq \lfloor \frac{c}{b} \rfloor, 0 \leq q \leq c \\ 0 & \text{otherwise} \end{cases}$$

for $0 \leq i \leq c-1$

$$M_0(p, q) = \begin{cases} p\mu_1/b & \text{if } q = p, 0 \leq p, q \leq c \\ (c - p)\mu_2 & \text{if } q = p+1, \text{rem}(p, b) = 0, 0 \leq p, q \leq c \\ 0 & \text{otherwise} \end{cases}$$

$$M_i(p, q) = \begin{cases} p\mu_1/b & \text{if } q = p-i, 0 \leq p, q \leq c, \text{rem}(q, b) = 0 \\ (c - p)\mu_2 & \text{if } i = 1, q = p, \text{rem}(q, b) = 0, 0 \leq p, q \leq c \\ 0 & \text{otherwise} \end{cases}$$

for $1 \leq i \leq b$

(D.5)

$$\begin{aligned}
X_i(p, q) &= \begin{cases} p\mu_1/b & \text{if } q = \frac{p-1}{b}, 0 \leq p \leq c, 0 \leq q \leq \lfloor \frac{c-1}{b} \rfloor \\ 0 & \text{otherwise} \end{cases} \\
&\text{for } 2 \leq i \leq b \\
X'_i(p, q) &= \begin{cases} p\mu_1/b & \text{if } q = \frac{p-1}{b}, 0 \leq p \leq c, 0 \leq q \leq \lfloor \frac{c-1}{b} \rfloor - 1 \\ 0 & \text{otherwise} \end{cases} \\
&\text{for } \max(2, \text{rem}(c, b) + 1) \leq i \leq b \\
L_i(p, q) &= \begin{cases} \lambda_i & \text{if } q = p, 0 \leq p, q \leq c \\ 0 & \text{otherwise} \end{cases} \\
&\text{for } i = 1, 2 \\
D_i(p, q) &= \begin{cases} \lambda_1 + \lambda_2 + p\mu_1 + (i - bp)\mu_2 & \text{if } q = p, 0 \leq p, q \leq \lfloor \frac{c}{b} \rfloor \\ 0 & \text{otherwise} \end{cases} \\
&\text{for } 0 \leq i \leq c - 1 \\
D_c(p, q) &= \begin{cases} \lambda_1 + \lambda_2 + p\mu_1/b + (c - p)\mu_2 & \text{if } q = p, 0 \leq p, q \leq c \\ -(c - p)\mu_2 & \text{if } q = p + 1, \text{rem}(p, b) \neq 0 \\ 0 & \text{otherwise} \end{cases}
\end{aligned} \tag{D.6}$$

The blocks H_m , $0 \leq m \leq c - 1$ and $H_{i,j}$, $i, j \geq 0$ are combined to form "super-blocks" which are defined as follows :

$$I_{-1} = \bigcup_{i=0}^{c-1} H_i, \text{ and } I_i = \bigcup_{j=0}^{\infty} H_{i,j}, i \geq 0. \tag{D.7}$$

The states in I_{-1} correspond to the existence of idle servers and the states in I_i , $i \geq 0$ correspond to those with exactly i Type 1-SR awaiting service. The matrices of

transition rates between and within these superblocks are denoted as :

$$\begin{aligned}
 B_{-1,-1} &: I_{-1} \rightarrow I_{-1} \\
 B_{-1,0} &: I_{-1} \rightarrow I_0 \\
 B_{0,-1} &: I_0 \rightarrow I_{-1} \\
 B_{0,0} &: I_0 \rightarrow I_0 \\
 A_0 &: I_i \rightarrow I_{i+1}, i \geq 0 \\
 A_1 &: I_i \rightarrow I_i, i \geq 1 \\
 A_2 &: I_i \rightarrow I_{i-1}, i \geq 1
 \end{aligned} \tag{D.8}$$

The aforementioned matrices of transition rates are partitioned into submatrices corresponding to transitions between the sub-blocks of I_i , $i \geq -1$. The entry to each of the aforementioned matrices is described below :

$$B_{-i,-1}(p,q) = \begin{cases} -D_p & \text{if } q = p \\ K_p & \text{if } q = p+1, b \neq 1 \\ J_p & \text{if } q = p-1, b \neq 1 \\ E_p & \text{if } q = p+b, b \neq 1 \\ F_p & \text{if } q = p-b, b \neq 1 \\ K_p + E_p & \text{if } q = p+1, b = 1 \\ J_p + F_p & \text{if } q = p-1, b = 1 \\ 0 & \text{otherwise} \end{cases} \tag{D.9}$$

for $0 \leq p, q \leq c-1$

$$\begin{aligned}
B_{-1,0}(p,q) &= \begin{cases} E_p & \text{if } q = 0, c-b \leq p \leq c-2 \\ E_{c-1} + K_{c-1} & \text{if } q = 0, p = c-1 \end{cases} \\
&\quad \text{for } 0 \leq p \leq c-1, 0 \leq q \leq \infty \\
B_{0,-1}(p,q) &= \begin{cases} J_c & \text{if } q = c-1, p = 0 \\ Fc & \text{if } q = c-b, p = 0, b \neq 1 \\ X_i & \text{if } q = c-i+p, 2 \leq i \leq b, 0 \leq p \leq b-1, \\ & q \geq [c - \text{rem}(c, b)], c-b+1 \leq q \leq c-1, \\ & \text{not}(i = b \text{ and } p = 0) \\ X'_i & \text{if } q = c-i+p, \max(2, \text{rem}(c, b) + 1) \leq i \leq b, \\ & 0 \leq p \leq b-1, q < [c - \text{rem}(c, b)], \\ & c-b+1 \leq q \leq c-1, \text{not}(i = b \text{ and } p = 0) \\ 0 & \text{otherwise} \end{cases} \\
&\quad \text{for } 0 \leq p \leq \infty, 0 \leq q \leq c-1 \\
B_{0,0}(p,q) &= \begin{cases} -D_c & \text{if } q = p \\ L_2 & \text{if } q = p+1 \\ M_i & \text{if } q = p-i, 1 \leq i \leq b \\ 0 & \text{otherwise} \end{cases} \\
&\quad \text{for } 0 \leq p, q \leq \infty \\
A_0(p,q) &= \begin{cases} L_1 & \text{if } q = p \\ 0 & \text{otherwise} \end{cases} \\
&\quad \text{for } 0 \leq p, q \leq \infty \\
A_1(p,q) &= \begin{cases} L_2 & \text{if } q = p+1 \\ -D_c & \text{if } q = p \\ 0 & \text{otherwise} \end{cases} \\
&\quad \text{for } 0 \leq p, q \leq \infty \\
A_2(p,q) &= \begin{cases} M_0 & \text{if } q = p \\ 0 & \text{otherwise} \end{cases} \\
&\quad \text{for } 0 \leq p, q \leq \infty
\end{aligned} \tag{D.10}$$

Given the block structure and the notations adopted, the transition rate matrix for this M/M/c system has the form

$$\mathbf{Q} = \begin{bmatrix} B_{-1,-1} & B_{-1,0} & 0 & 0 & 0 & \cdots \\ B_{0,-1} & B_{0,0} & A_0 & 0 & 0 & \cdots \\ 0 & A_2 & A_1 & A_0 & 0 & \cdots \\ 0 & 0 & A_2 & A_1 & A_0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \quad (\text{D.11})$$

Invariant Measures

The process described by the above transition rate matrix is the Markov Process of Quasi-GI/M/1 Type with Complex Boundary Behaviour (QGM1-CB process). The steady state probability vector of $\mathbf{Q}$ is of matrix geometric form where

$$\underline{\pi} = (\underline{\pi}_{-1}, \underline{\pi}_0, \underline{\pi}_0 R, \underline{\pi}_0 R^2, \underline{\pi}_0 R^3, \dots) \quad (\text{D.12})$$

and satisfies

$$\underline{\pi} \mathbf{Q} = \underline{0}, \quad \underline{\pi} \underline{e} = 1 \quad (\text{D.13})$$

where $\underline{0}$ is a zero-row vector of infinite size and $\underline{e}$ is column vector of ones of appropriate size. The rate matrix R is the minimal solution of

$$A_0 + RA_1 + R^2 A_2 = \mathbf{0}. \quad (\text{D.14})$$

and has the following structure [21, 20] :

$$R = \begin{bmatrix} R_0 & R_1 & R_2 & \cdots \\ 0 & R_0 & R_1 & \cdots \\ 0 & 0 & R_0 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \quad (\text{D.15})$$

where the submatrices R_i 's, $i \geq 0$ are matrices of dimension $(c+1) \times (c+1)$.

Substitution of Eqn.(D.15) into Eqn.(D.14) yields the following set of equations for solving the R_i 's, $i \geq 0$:

$$\begin{aligned} L_1 - R_0 D_c + R_0^2 M_0 &= 0 \\ R_{i-1} L_2 - R_i D_c + \sum_{j=0}^i R_j R_{i-j} M_0 &= 0, \quad i \geq 1 \end{aligned} \quad (D.16)$$

The above set of equations for R_i 's ($i \geq 0$) can be solved numerically by transforming the system to the equivalent system of embedded Markov chain where the system is observed at the state transitions.

Let

$$\tilde{L}_1 = D_c^{-1} L_1, \quad \tilde{L}_2 = D_c^{-1} L_2, \quad \tilde{M}_0 = D_c^{-1} M_0, \quad \text{and} \quad \tilde{R}_0 = D_c^{-1} R_0 D_c, \quad (D.17)$$

then Eqn.(D.16) becomes

$$\begin{aligned} \tilde{R}_0 &= \tilde{L}_1 + \tilde{R}_0^2 \tilde{M}_0 \\ \tilde{R}_i &= \tilde{R}_{i-1} \tilde{L}_2 + \sum_{j=0}^{i-1} \tilde{R}_j \tilde{R}_{i-j} \tilde{M}_0 + (\tilde{R}_0 \tilde{R}_i + \tilde{R}_i \tilde{R}_0) \tilde{M}_0, \quad i \geq 1 \end{aligned} \quad (D.18)$$

The first equation is used for solving $\tilde{R}_0$ by iteration. Initial $\tilde{R}_0$ is a zero matrix of size $(c+1) \times (c+1)$. Iterations can be terminated when the maximum element-wise difference between the successive $\tilde{R}_0$'s is less than 10^{-11} . Using the same iteration approach, $\tilde{R}_i$'s, $i > 0$ can be solved using the solution for $\tilde{R}_j$, $j < i$. In the study, the computation of $\tilde{R}_i$ is stopped when the iterative procedure stops in the first iteration and when $(i+1)$ is a multiple of b . The desired R_i 's are obtained by backward transformation :

$$R_i = D_c \tilde{R}_i D_c^{-1}, \quad i \geq 0. \quad (D.19)$$

Internal accuracy check is performed to verify the accuracy of the computation of R (or $\tilde{R}$). Let

$$\tilde{R}^* = \sum_{i=0}^{\infty} \tilde{R}_i, \quad (D.20)$$

then from Eqn.(D.18), it follows that

$$\tilde{R}^* = \tilde{L}_1 + \tilde{R}^* \tilde{L}_2 + \left(\tilde{R}^* \right)^2 \tilde{M}_0. \quad (D.21)$$

Iterative method with initial $\tilde{R}^* =$ zero matrix of size $(c+1) \times (c+1)$ is then used. The computation of the rate matrix R is considered accurate when largest element-wise difference between the $\tilde{R}^*$ obtained from Eqn.(D.20) and Eqn.(D.21) is less than 10^{-8} .

M/G/1 Paradigm and Invariant Probability Vector

According to the theory of matrix-geometric invariant vector outlined in chapter 2, for the system which satisfies

$$\underline{\pi} \mathbf{Q} = \underline{0}, \quad \underline{\pi} \underline{e} = 1, \quad (D.22)$$

then the invariant probability vector $\underline{\pi}$ will have the form

$$\underline{\pi} = (\underline{\pi}_{-1}, \underline{\pi}_0, \underline{\pi}_1, \underline{\pi}_2, \dots) \quad (D.23)$$

where $\underline{\pi}_i$ corresponds to the probability vector on the super-block I_i , $i \geq -1$ and

$$\underline{\pi}_{i+1} = \underline{\pi}_i R, \quad i \geq 0. \quad (D.24)$$

The boundary probability vector $\underline{\pi}_{-1}$ and $\underline{\pi}_0$ can be obtained first by solving the following subsystem :

$$(\underline{\pi}_{-1} \quad \underline{\pi}_0) Q_{I_{-1} \cup I_0} = (\underline{0} \quad \underline{0}) \quad (D.25)$$

where

$$Q_{I_{-1} \cup I_0} = \begin{bmatrix} B_{-1,-1} & B_{-1,0} \\ B_{0,-1} & B_{0,0} + RA_2 \end{bmatrix} \quad (D.26)$$

and

$$\underline{\pi}_{-1} = (\underline{\pi}_{-1,0}, \underline{\pi}_{-1,1}, \dots, \underline{\pi}_{-1,c-1})$$

$$\underline{\pi}_0 = (\underline{\pi}_{0,0}, \underline{\pi}_{0,1}, \underline{\pi}_{0,2}, \dots)$$

$$(D.27)$$

$$\underline{\pi}_{-1,i} = \text{Prob. vector on states in block } H_i, \quad 0 \leq i \leq c-1$$

$$\underline{\pi}_{0,i} = \text{Prob. vector on states in block } H_{0,i}, \quad i \geq 0$$

Note : $Q_{I_{-1} \cup I_0}$ is a matrix with negative diagonal entries, non-negative off-diagonal entries, and row sums equal to 0; hence it can be treated as a transition rate matrix for the Markov process on $I_{-1} \cup I_0$.

The matrix $(B_{0,0} + RA_2)$ has the following structure :

$$(B_{0,0} + RA_2) = \begin{bmatrix} C_1 & C_2 & C_3 & C_4 & \cdots \\ C_0 & C_1 & C_2 & C_3 & \cdots \\ 0 & C_0 & C_1 & C_2 & \cdots \\ 0 & 0 & C_0 & C_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \quad (D.28)$$

Each C_i 's, $i \geq 0$ is a square matrix of order $(b \times (c+1))$, formed by blocks of submatrices of size $(c+1) \times (c+1)$:

$$\begin{aligned} C_0 &= \begin{bmatrix} M_b & M_{b-1} & M_{b-2} & \cdots & M_1 \\ 0 & M_b & M_{b-1} & \cdots & M_2 \\ 0 & 0 & M_b & \cdots & M_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & M_b \end{bmatrix} \\ C_1 &= \begin{bmatrix} -D_c + R_0 M_0 & L_2 + R_1 M_0 & R_2 M_0 & \cdots & R_{b-1} M_0 \\ M_1 & -D_c + R_0 M_0 & L_2 + R_1 M_0 & \cdots & R_{b-2} M_0 \\ M_2 & M_1 & -D_c + R_0 M_0 & \cdots & R_{b-2} M_0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ M_{b-1} & M_{b-2} & M_{b-3} & \cdots & -D_c + R_0 M_0 \end{bmatrix} \\ C_2 &= \begin{bmatrix} R_b M_0 & R_{b+1} M_0 & R_{b+2} M_0 & \cdots & R_{2b-1} M_0 \\ R_{b-1} M_0 & R_b M_0 & R_{b+1} M_0 & \cdots & R_{2b-2} M_0 \\ R_{b-2} M_0 & R_{b-1} M_0 & R_b M_0 & \cdots & R_{2b-3} M_0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ L_2 + R_1 M_0 & R_2 M_0 & R_3 M_0 & \cdots & R_b M_0 \end{bmatrix} \end{aligned} \quad (D.29)$$

$$C_i = \begin{bmatrix} R_{(i-1)b}M_0 & R_{(i-1)b+1}M_0 & R_{(i-1)b+2}M_0 & \cdots & R_{ib-1}M_0 \\ R_{(i-1)b-1}M_0 & R_{(i-1)b}M_0 & R_{(i-1)b+1}M_0 & \cdots & R_{ib-2}M_0 \\ R_{(i-1)b-2}M_0 & R_{(i-1)b-1}M_0 & R_{(i-1)b}M_0 & \cdots & R_{ib-3}M_0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R_{(i-2)b+1}M_0 & R_{(i-2)b+2}M_0 & R_{(i-2)b+3}M_0 & \cdots & R_{(i-1)b}M_0 \end{bmatrix}, i \geq 3 \quad (D.30)$$

The matrix $(B_{0,0} + RA_2)$ has M/G/1 paradigm. In order to solve the boundary vector $(\pi_{-1} \pi_0)$, the Markov process on $I_{-1} \cup I_0 \left(= \left(\bigcup_{m=0}^{c-1} H_m \right) \cup \left(\bigcup_{j=0}^{\infty} H_{0,j} \right) \right)$ is converted to an embedded Markov chain on $I_{-1} \cup I_0$ and the following boundary vector (with finite size) can be determined first :

$$(\pi_{-1}, \pi_{0,0}, \pi_{0,1}, \dots, \pi_{0,b-1}) \quad (D.31)$$

where

$$(\pi_{-1}, \pi_{0,0}, \pi_{0,1}, \dots, \pi_{0,b-1}) \tilde{Q} = \underline{0} \quad (D.32)$$

The transformation from the Markov process to the embedded Markov chain on $I_{-1} \cup I_0$ can be done by letting :

$$\begin{aligned} \tilde{C}_1 &= D_{gc}^{-1} C_1 + I, \text{ } I \text{ is identity matrix of order } (c+1) * b \\ \tilde{C}_i &= D_{gc}^{-1} C_i, \text{ } i = 0, 2, 3, 4, \dots \\ \tilde{K}_i &= D_i^{-1} K_i, \text{ } 0 \leq i \leq c \\ \tilde{J}_i &= D_i^{-1} J_i, \text{ } 0 \leq i \leq c \\ \tilde{F}_i &= D_i^{-1} F_i, \text{ } 0 \leq i \leq c \\ \tilde{E}_i &= D_i^{-1} E_i, \text{ } 0 \leq i \leq c \\ \tilde{X}_i &= D_i^{-1} X_i, \text{ } 2 \leq i \leq b \\ \tilde{X}'_i &= D_i^{-1} X'_i, \text{ } \max(2, rem+1) \leq i \leq b \end{aligned} \quad (D.33)$$

where

$$D_{gc}(p, q) = \begin{cases} D_c & \text{if } q = p \\ 0 & \text{otherwise} \end{cases} \quad \text{for } 0 \leq p, q \leq b-1. \quad (\text{D.34})$$

Following Lucantoni's [2] and Miller's [20] approach, the following matrices are determined first :

1. G : a square matrix of order $(c+1) \times b$ denoting the hitting probabilities of states in $\bigcup_{n=0}^{b-1} H_{0,jb+n}$ starting at $\bigcup_{n=0}^{b-1} H_{0,(j+1)b+n}$, $j \geq 0$ with G satisfies

$$G = \sum_{v=0}^{\infty} \tilde{C}_v G^v. \quad (\text{D.35})$$

G can be found iteratively following Step 3 in Appendix A.

2. V : a square matrix of order $(c+1) \times b$ denoting the first passage probabilities of hitting states in $\bigcup_{n=0}^{b-1} H_{0,(j+1)b+n}$ starting from states in $\bigcup_{n=0}^{b-1} H_{0,jb+n}$ without first return to $\bigcup_{n=0}^{b-1} H_{0,jb+n}$, $j \geq 0$. V satisfies

$$V = \sum_{v=0}^{\infty} \tilde{C}_v G^{v-2}. \quad (\text{D.36})$$

3. W : a square matrix of order $(c+1) \times b$ denoting the first passage probabilities of hitting states in $\bigcup_{n=0}^{b-1} H_{0,(j+1)b+n}$ starting from states in $\bigcup_{n=0}^{b-1} H_{0,(j+1)b+n}$ without visiting to $\bigcup_{n=0}^{b-1} H_{0,jb+n}$, $j \geq 0$. W satisfies

$$W = \sum_{v=1}^{\infty} \tilde{C}_v G^{v-1}. \quad (\text{D.37})$$

4. $\tilde{S}_i$, $i \geq 1$: a square matrix of order $(c+1) \times b$ whose entries are the expected number of visits to states in $\bigcup_{n=0}^{b-1} H_{0,(j+1)b+n}$ starting from states in $\bigcup_{n=0}^{b-1} H_{0,jb+n}$ before hitting

the states in $\bigcup_{n=0}^{b-1} H_{0,kb+n}$, $k < j+i$. $\tilde{S}_i$, $i \geq 1$ is computed as

$$\tilde{S}_i = \sum_{v=i+1}^{\infty} \tilde{C}_v G^{v-i+1} (I - W)^{-1} \quad (D.38)$$

where I is an identity matrix of the same order.

After the transformation from the Markov process to the embedded Markov chain, the $\tilde{Q}$ matrix, which describes the transitions between the set of states in blocks $\left(\bigcup_{m=0}^{c-1} H_m\right) \cup \left(\bigcup_{j=0}^{b-1} H_{0,j}\right)$ and is shown in section 4.3, is transformed to become the transition probability matrix $\tilde{P}$ which has the following structure :

$$\tilde{P} = \begin{bmatrix} 0 & \tilde{K}_0 & 0 & & E_0 & 0 & & \\ \tilde{J}_1 & 0 & \tilde{K}_1 & \dots & 0 & E_1 & & \\ 0 & \tilde{J}_2 & 0 & & 0 & 0 & \ddots & \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \\ \tilde{F}_b & 0 & 0 & & 0 & \tilde{K}_b & & \\ 0 & \tilde{F}_{b+1} & 0 & & \tilde{J}_{b+1} & 0 & & \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \ddots & \\ & & & & \ddots & \vdots & \vdots & \\ & & & & \tilde{E}_{c-b-1} & 0 & \dots & \\ & & & & 0 & \tilde{E}_{c-b} & \dots & \\ & & & & \vdots & \vdots & \ddots & \\ & & & & \ddots & \tilde{K}_{c-2} & \tilde{E}_{c-2} & 0 & \ddots & \vdots \\ \tilde{F}_{c-1} & 0 & \tilde{J}_{c-1} & 0 & \tilde{E}_{c-1} + \tilde{K}_{c-1} & 0 & \dots & 0 \\ \hline & & \tilde{F}_c & \dots & \tilde{X}_2 & \tilde{J}_c & & \\ & & \ddots & \tilde{X}_3 & \tilde{X}_2 & & & \\ & & & \vdots & \vdots & & \tilde{C}_1 + \tilde{S}_1 \tilde{C}_0 & \\ & & & 0 & \tilde{X}_b & & & 0 \end{bmatrix} \quad (D.39)$$

The invariant probability vector of $\tilde{\mathbf{P}}$ is denoted as

$$\begin{aligned} &(\tilde{z}_{-1}, \tilde{z}_{0,0}, \tilde{z}_{0,1}, \dots, \tilde{z}_{0,b-1}) \text{ and} \\ &\tilde{z}_{-1} = (\tilde{z}_{-1,0}, \tilde{z}_{-1,1}, \tilde{z}_{-1,2}, \dots, \tilde{z}_{-1,c-1}) \end{aligned} \quad (\text{D.40})$$

which can be solved by iteration according to

$$(\tilde{z}_{-1}, \tilde{z}_{0,0}, \tilde{z}_{0,1}, \dots, \tilde{z}_{0,b-1}) = (\tilde{z}_{-1}, \tilde{z}_{0,0}, \tilde{z}_{0,1}, \dots, \tilde{z}_{0,b-1}) \tilde{\mathbf{P}}. \quad (\text{D.41})$$

Since $(\tilde{z}_{-1}, \tilde{z}_{0,0}, \tilde{z}_{0,1}, \dots, \tilde{z}_{0,b-1})$ is the invariant probability vector of $\tilde{\mathbf{P}}$, it is then converted back to the invariant probability of $\mathbf{Q}$ as

$$\begin{aligned} \underline{y}_{-1,i} &= \tilde{z}_{-1,i} D_i^{-1}, \quad 0 \leq i \leq c-1 \text{ and} \\ (\underline{y}_{0,0}, \underline{y}_{0,1}, \dots, \underline{y}_{0,b-1}) &= (\tilde{z}_{0,0}, \tilde{z}_{0,1}, \dots, \tilde{z}_{0,b-1}) D_{gc}^{-1}. \end{aligned} \quad (\text{D.42})$$

The actual invariant probability vectors on $I_{-1} \cup \left(\bigcup_{i=0}^{b-1} H_{0,i} \right)$ are computed as

$$\begin{aligned} \pi_{-1,i} &= \left(\frac{1-\rho}{\xi} \right) \underline{y}_{-1,i}, \quad 0 \leq i \leq c-1, \quad \text{and} \\ (\pi_{0,0}, \pi_{0,1}, \pi_{0,2}, \dots, \pi_{0,b-1}) &= \left(\frac{1-\rho}{\xi} \right) (\underline{y}_{0,0}, \underline{y}_{0,1}, \underline{y}_{0,2}, \dots, \underline{y}_{0,b-1}) \end{aligned} \quad (\text{D.43})$$

where

$$\begin{aligned} \rho &= \text{system utilization} = \frac{b\lambda_1}{c\mu_1} + \frac{\lambda_2}{c\mu_2}, \quad \text{and} \\ \xi &= \sum_{i=0}^{c-1} \frac{c-i}{c} \underline{y}_{-1,i} \underline{e}, \\ \underline{e} &= \text{unit column vector of size } \left[\frac{i}{b} \right]. \end{aligned} \quad (\text{D.44})$$

Then the rest of the probability vectors in super-state $I_0, \pi_{0,b}, \pi_{0,b+1}, \pi_{0,b+2}, \dots, \pi_{0,\infty}$ can be obtained as

$$\begin{aligned} (\pi_{0,(i+1)b}, \pi_{0,(i+1)b+1}, \dots, \pi_{0,(i+2)b-1}) &= \sum_{j=0}^i (\pi_{0,jb}, \pi_{0,jb+1}, \dots, \pi_{0,(j+1)b-1}) \\ &\quad D_{gc} \tilde{S}_{i+1-j} D_{gc}^{-1}, \quad i \geq 0. \end{aligned} \quad (\text{D.45})$$

For $\pi_{i,j}, i, j \geq 0$, the entries of the rate matrix R are employed for the computation using the following equations :

$$\pi_{i+1,j} = \sum_{k=0}^j \pi_{i,k} R_{j-k}. \quad (\text{D.46})$$

The computation of the invariant probability vector can be truncated to any level. In the implementation, the truncation level is determined when $\sum_{j=0}^{\infty} \pi_{i+1,j} e < 10^{-9}$ is satisfied.

Appendix E

Algorithmic Solutions For Non-Preemptive Priority M/M/c Model (Analysis 2 — Sect. 4.4)

Following is a variation of the methodology presented in Appendix D to solve the M/M/c queueing model where Type 2-SR has absolute non-preemptive priority over Type 1-SR. The procedures for solving the invariant probability vector is given below.

The state space of the system, which is the same as in section 4.3, is repeated here :

1. Let $x_{i,j,k}$ be the state with i Type 1-SR waiting in the Type 1 infinite queue, j Type 2-SR waiting in the Type 2 infinite queue, and k servers serving Type 1-SRs and $(c - k)$ servers serving Type 2-SRs, $i, j \geq 0$, $0 \leq k \leq c$. Then define the blocks as

$$H_{i,j} = \{x_{i,j,k} \mid 0 \leq k \leq c\}, \quad i, j \geq 0. \quad (\text{E.1})$$

2. Let $x_{m,n}$ be the state with m server being in service, n Type 1-SR being serviced and no SR is waiting in any of Type 1 or Type 2 queue, $0 \leq m \leq (c - 1)$, $0 \leq n \leq \lfloor \frac{m}{b} \rfloor$, where $\lfloor y \rfloor$ is the largest integer no greater than y . Then define the blocks as

$$H_m = \left\{ x_{m,n} \mid 0 \leq n \leq \left\lfloor \frac{m}{b} \right\rfloor \right\}, \quad 0 \leq m \leq (c - 1). \quad (\text{E.2})$$

The state space of the system is then given as

$$S = \left(\bigcup_{m=0}^{c-1} H_m \right) \cup \left(\bigcup_{i \geq 0} \bigcup_{j \geq 0} H_{i,j} \right). \quad (\text{E.3})$$

The matrices of transition rates between the blocks defined above are denoted as :

$$\begin{aligned} K_i &: H_i \rightarrow H_{i+1}, 0 \leq i \leq c-1 \\ J_i &: H_i \rightarrow H_{i-1}, 1 \leq i \leq c-1 \\ J_c &: H_{0,0} \rightarrow H_{c-1} \\ F_i &: H_i \rightarrow H_{i-b}, b \leq i \leq c-1 \\ F_c &: H_{0,0} \rightarrow H_{c-b} \\ E_i &: H_i \rightarrow H_{i+b}, 0 \leq i \leq c-b-1 \\ E_j &: H_j \rightarrow H_{0,0}, c-b \leq j \leq c-1 \\ M_0 &: H_{i,0} \rightarrow H_{i-1,0}, i \geq 1 \\ M_i &: H_{k,j} \rightarrow H_{k,j-i}, k \geq 0, 1 \leq i \leq b, j \geq 1, j-i \geq 0 \\ X_i &: H_{0,j} \rightarrow H_{c-i+j}, 2 \leq i \leq b, 0 \leq j \leq b-1, c-i+j \geq (c - \text{rem}(c, b)) \\ X'_i &: H_{0,j} \rightarrow H_{c-i+j}, 2 \leq i \leq b, 0 \leq j \leq b-1, c-i+j < (c - \text{rem}(c, b)) \\ Z_j &: H_{i,j} \rightarrow H_{i-1,0}, i \geq 1, 1 \leq j \leq b \\ L_1 &: H_{i,j} \rightarrow H_{i+1,j}, i, j \geq 0 \\ L_2 &: H_{i,j} \rightarrow H_{i,j+1}, i, j \geq 0 \\ -D_i &: H_i \rightarrow H_i, 0 \leq i \leq c-1 \\ -D'_c &: H_{i,0} \rightarrow H_{i,0}, i \geq 0 \\ -D_c &: H_{i,j} \rightarrow H_{i,j}, i \geq 0, j \geq 1 \end{aligned} \quad (\text{E.4})$$

* $\text{rem}(c, b)$ is the remainder of the division : c/b .

The entry to each of the matrices mentioned above is given below :

$$\begin{aligned}
 K_i(p, q) &= \begin{cases} \lambda_2 & \text{if } q = p, i \neq c-1, 0 \leq p \leq \lfloor \frac{c}{b} \rfloor, 0 \leq q \leq \lfloor \frac{c-1}{b} \rfloor \\ \lambda_2 & \text{if } q = pb, i = c-1, 0 \leq p \leq \lfloor \frac{c}{b} \rfloor, 0 \leq q \leq c \\ 0 & \text{otherwise} \end{cases} \\
 &\text{for } 0 \leq i \leq c-1 \\
 J_i(p, q) &= \begin{cases} (i-bp)\mu_2 & \text{if } q = p, i \neq c, 0 \leq p \leq \lfloor \frac{c}{b} \rfloor, 0 \leq q \leq \lfloor \frac{c-1}{b} \rfloor \\ (c-p)\mu_2 & \text{if } q = \frac{p}{b}, i = c, 0 \leq p \leq c, 0 \leq q \leq \lfloor \frac{c-1}{b} \rfloor \\ p\mu_1/b & \text{if } q = \frac{p-1}{b}, i = c, 0 \leq p \leq c, 0 \leq q \leq \lfloor \frac{c-1}{b} \rfloor \\ 0 & \text{otherwise} \end{cases} \\
 &\text{for } 1 \leq i \leq c \\
 F_i(p, q) &= \begin{cases} p\mu_1 & \text{if } q = p-1, i \neq c, 0 \leq p \leq \lfloor \frac{c}{b} \rfloor, 0 \leq q \leq \lfloor \frac{c}{b} \rfloor - 1 \\ p\mu_1/b & \text{if } q = \frac{p}{b} - 1, b \neq 1, i = c, 0 \leq p \leq c, 0 \leq q \leq \lfloor \frac{c}{b} \rfloor - 1 \\ 0 & \text{otherwise} \end{cases} \\
 &\text{for } b \leq i \leq c \\
 E_i(p, q) &= \begin{cases} \lambda_1 & \text{if } q = p+1, i < c-b, 0 \leq p \leq \lfloor \frac{c}{b} \rfloor, 0 \leq q \leq \lfloor \frac{c}{b} \rfloor + 1 \\ \lambda_1 & \text{if } q = pb + c - i, i \geq c-b, 0 \leq p \leq \lfloor \frac{c}{b} \rfloor, 0 \leq q \leq c \\ 0 & \text{otherwise} \end{cases} \\
 &\text{for } 0 \leq i \leq c-1 \\
 M_0(p, q) &= \begin{cases} p\mu_1/b & \text{if } q = p, 0 \leq p, q \leq c \\ (c-p)\mu_2 & \text{if } q = p+1, \text{rem}(p, b) = 0, 0 \leq p, q \leq c \\ 0 & \text{otherwise} \end{cases} \\
 M_i(p, q) &= \begin{cases} p\mu_1/b & \text{if } q = p-i, 0 \leq p, q \leq c, \text{rem}(q, b) = 0 \\ (c-p)\mu_2 & \text{if } i = 1, q = p, 0 \leq p, q \leq c \\ 0 & \text{otherwise} \end{cases} \\
 &\text{for } 1 \leq i \leq b
 \end{aligned}$$

(E.5)

$$\begin{aligned}
X_i(p, q) &= \begin{cases} p\mu_1/b & \text{if } q = \frac{p-1}{b}, 0 \leq p \leq c, 0 \leq q \leq \lfloor \frac{c-1}{b} \rfloor \\ 0 & \text{otherwise} \end{cases} \\
&\text{for } 2 \leq i \leq b \\
X'_i(p, q) &= \begin{cases} p\mu_1/b & \text{if } q = \frac{p-1}{b}, 0 \leq p \leq c, 0 \leq q \leq \lfloor \frac{c-1}{b} \rfloor - 1 \\ 0 & \text{otherwise} \end{cases} \\
&\text{for } \max(2, \text{rem}(c, b) + 1) \leq i \leq b \\
Z_i(p, q) &= \begin{cases} p\mu_1/b & \text{if } q = p - i, 0 \leq p, q \leq c, 0 \leq \text{rem}(q, b) \leq b - i \\ 0 & \text{otherwise} \end{cases} \\
&\text{for } 1 \leq i \leq b - 1 \\
L_i(p, q) &= \begin{cases} \lambda_i & \text{if } q = p, 0 \leq p, q \leq c \\ 0 & \text{otherwise} \end{cases} \\
&\text{for } i = 1, 2 \\
D_i(p, q) &= \begin{cases} \lambda_1 + \lambda_2 + p\mu_1 + (i - bp)\mu_2 & \text{if } q = p, 0 \leq p, q \leq \lfloor \frac{c}{b} \rfloor, i \neq c \\ \lambda_1 + \lambda_2 + p\mu_1/b + (c - p)\mu_2 & \text{if } q = p, 0 \leq p, q \leq \lfloor \frac{c}{b} \rfloor, i = c \\ -(c - p)\mu_2 & \text{if } q = p - 1, \text{rem}(p, b) \neq 0, i = c \\ 0 & \text{otherwise} \end{cases} \\
&\text{for } 0 \leq i \leq c \\
D'_c(p, q) &= \begin{cases} \lambda_1 + \lambda_2 + p\mu_1/b + (c - p)\mu_2 & \text{if } q = p, 0 \leq p, q \leq c \\ 0 & \text{otherwise} \end{cases}
\end{aligned} \tag{E.6}$$

The blocks H_m , $0 \leq m \leq c - 1$ and $H_{i,j}$, $i, j \geq 0$ are combined to form “super-blocks” which are defined as follows :

$$I_{-1} = \bigcup_{i=0}^{c-1} H_i, \text{ and } I_j = \bigcup_{i=0}^{\infty} H_{i,j}, j \geq 0 \tag{E.7}$$

The states in I_{-1} are exactly those corresponding to the existence of idle servers. The states in I_i , $i \geq 0$ are those with exactly i Type 2-SR awaiting service. The matrices of

transition rates between and within these superblocks are denoted as :

$$\begin{aligned}
 B_{i,-1} &: I_i \rightarrow I_{-1}, \quad -1 \leq i \leq b-1 \\
 B_{i,0} &: I_i \rightarrow I_0, \quad -1 \leq i \leq b-1 \\
 A_j &: I_i \rightarrow I_{i-j+1}, \quad i \geq 0, \quad 0 \leq j \leq b+1, \quad i-j \geq 0 \\
 A_{b+1} &: I_b \rightarrow I_0
 \end{aligned} \tag{E.8}$$

The aforementioned matrices of transition rates are partitioned into submatrices corresponding to transitions between the sub-blocks of I_i , $i \geq -1$. The entry to each of the aforementioned matrices is described below :

$$B_{-1,-1}(p,q) = \begin{cases} -D_p & \text{if } q = p \\ K_p & \text{if } q = p+1, b \neq 1 \\ J_p & \text{if } q = p-1, b \neq 1 \\ E_p & \text{if } q = p+b, b \neq 1 \\ F_p & \text{if } q = p-b, b \neq 1 \\ K_p + E_p & \text{if } q = p+1, b = 1 \\ J_p + F_p & \text{if } q = p-1, b = 1 \\ 0 & \text{otherwise} \end{cases} \tag{E.9}$$

for $0 \leq p, q \leq c-1$

$$B_{-1,0}(p,q) = \begin{cases} E_p & \text{if } q = 0, c-b \leq p \leq c-2 \\ K_{c-1} + E_{c-1} & \text{if } q = 0, p = c-1 \\ 0 & \text{otherwise} \end{cases}$$

for $0 \leq p \leq c, 0 \leq q \leq \infty$

$$B_{k,-1}(p, q) = \begin{cases} J_c & \text{if } q = c - 1, p = 0 \\ Fc & \text{if } q = c - b, p = 0, b \neq 1 \\ X_i & \text{if } q = c - i + k, 2 \leq i \leq b, \text{ not}(i = b \text{ and } k = 0), \\ & q \geq [c - \text{rem}(c, b)], c - b + 1 \leq q \leq c - 1 \\ X'_i & \text{if } q = c - i + k, \max(2, \text{rem}(c, b) + 1) \leq i \leq b, \\ & \text{not}(i = b \text{ and } k = 0), q < [c - \text{rem}(c, b)], \\ & c - b + 1 \leq q \leq c - 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{for } 0 \leq p \leq \infty, 0 \leq q \leq c - 1, 0 \leq k \leq b - 1$$

$$B_{k,0}(p, q) = \begin{cases} -D_c & \text{if } q = p, k = 0 \\ L_1 & \text{if } q = p + 1, k = 0 \\ M_0 & \text{if } q = p - 1, k = 0 \\ M_i & \text{if } q = p, k \neq 0 \\ Z_k & \text{if } q = p + 1, k \neq 0 \\ 0 & \text{otherwise} \end{cases} \quad (\text{E.10})$$

$$\text{for } 0 \leq p, q \leq \infty, 0 \leq k \leq b - 1$$

$$A_0(p, q) = \begin{cases} L_2 & \text{if } q = p \\ 0 & \text{otherwise} \end{cases}$$

$$\text{for } 0 \leq p, q \leq \infty$$

$$A_1(p, q) = \begin{cases} L_1 & \text{if } q = p + 1 \\ -D'_c & \text{if } q = p \\ 0 & \text{otherwise} \end{cases}$$

$$\text{for } 0 \leq p, q \leq \infty$$

$$A_k(p, q) = \begin{cases} M_{k-1} & \text{if } q = p \\ 0 & \text{otherwise} \end{cases}$$

$$\text{for } 0 \leq p, q \leq \infty, 2 \leq k \leq b + 1$$

Given the block structure and the notations adopted, it is be shown clearly from Fig.4.7 that the transition rate matrix for this M/M/c system has the form

$$\mathbf{Q} = \begin{bmatrix} B_{-1,-1} & B_{-1,0} & 0 & 0 & 0 & \cdots \\ B_{0,-1} & B_{0,0} & A_0 & 0 & 0 & \cdots \\ B_{1,-1} & B_{1,0} & A_1 & A_0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \\ B_{b-1,-1} & B_{b-1,0} & A_{b-1} & A_{b-2} & A_{b-3} & \cdots \\ 0 & A_{b+1} & A_b & A_{b-1} & A_{b-2} & \cdots \\ 0 & 0 & A_{b+1} & A_b & A_{b-1} & \cdots \\ 0 & 0 & 0 & A_{b+1} & A_b & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \quad (\text{E.11})$$

Invariant Measures

The process described by the above transition rate matrix is Markov Process of Quasi-GI/M/1 Type with Complex Boundary Behaviour (QGM1-CB process). The steady state probability vector of $\mathbf{Q}$ is of matrix geometric form where

$$\underline{\pi} = (\pi_{-1}, \pi_0, \pi_1, \pi_2, \pi_3, \dots) = (\pi_{-1}, \pi_0, \pi_0 R, \pi_0 R^2, \pi_0 R^3, \dots) \quad (\text{E.12})$$

and satisfies

$$\underline{\pi} \mathbf{Q} = \underline{0}, \quad \underline{\pi} \underline{e} = 1 \quad (\text{E.13})$$

where $\underline{0}$ is a zero-row vector of infinite size and $\underline{e}$ is column vector of ones of appropriate size. The rate matrix R is called the "Rate matrix" of the QGM1-CB process and it is the minimal solution of

$$A_0 + RA_1 + R^2 A_2 + \cdots + R^{b+1} A_{b+1} = 0. \quad (\text{E.14})$$

The definition of R is similar to that given in Analysis 1 and the matrix R again has the same Toeplitz structure [21, 20] :

$$R = \begin{bmatrix} R_0 & R_1 & R_2 & \cdots \\ 0 & R_0 & R_1 & \cdots \\ 0 & 0 & R_0 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \quad (\text{E.15})$$

where the submatrices R_i 's, $i \geq 0$ are matrices of dimension $(c+1) \times (c+1)$.

Substitution of Eqn.(E.15) into Eqn.(E.14) yields the following set of equations for solving the R_i 's, $i \geq 0$:

$$\begin{aligned} L_2 - R_{1,0}D'_c + \sum_{i=2}^{b+1} R_{i,0}M_{i-1} &= 0 \\ R_{i,i-1}L_1 - R_{1,i}D'_c + \sum_{j=2}^{b+1} R_{j,i}M_{j-1} &= 0, \quad i \geq 1 \end{aligned} \quad (\text{E.16})$$

where

$$R_{j+1,i} \triangleq i^{th} \text{ submatrix of } R^{j+1} = \sum_{k=0}^i R_{1,k}R_{j,i-k}, \quad i \geq 0, \quad j > 0. \quad (\text{E.17})$$

Note : Since R is a upper Toeplitz matrix, R^k is also upper Toeplitz matrix.

The above set of equations for R_i 's ($i \geq 0$) can be solved numerically by transforming the system to the equivalent system of embedded Markov chain where the system is observed at the state transitions.

Let

$$\begin{aligned} \tilde{L}_1 &= (D'_c)^{-1} L_1, \\ \tilde{L}_2 &= (D'_c)^{-1} L_2, \\ \tilde{M}_i &= (D'_c)^{-1} M_i, \quad 1 \leq i \leq b, \text{ and} \\ \tilde{R}_0 &= (D'_c)^{-1} R_0 D'_c, \end{aligned} \quad (\text{E.18})$$

and due to the special structure of R , therefore $(R^i)_0 = (R_0)^i$, and hence

$$(\tilde{R}_0)^i = \tilde{R}_{i,0} = (D'_c)^{-1} (R_0)^i D'_c. \quad (\text{E.19})$$

Eqn.(E.16) then becomes

$$\tilde{R}_{1,0} = \tilde{L}_2 + \sum_{i=2}^{b+1} \tilde{R}_{i,0} \tilde{M}_{i-1} \quad (\text{E.20a})$$

$$\tilde{R}_{1,i} = \tilde{R}_{1,i-1} \tilde{L}_1 + \sum_{j=1}^b \tilde{R}_{j+1,i} \tilde{M}_j, \quad i \geq 1 \quad (\text{E.20b})$$

The first equation is used for solving $\tilde{R}_{1,0}$ by iteration. The following steps are to be followed :

- i. Let initial $\tilde{R}_{1,0} = \tilde{L}_2$.
- ii. Compute $\tilde{R}_{2,0}, \tilde{R}_{3,0}, \dots, \tilde{R}_{b+1,0}, \left(\tilde{R}_{i,0} = (\tilde{R}_0)^i, i = 2, 3, \dots, b+1 \right)$.
- iii. Compute the new $\tilde{R}_{1,0}$ using Eqn.(E.20a).
- iv. Repeat step (ii) until maximum element-wise difference between the successive $\tilde{R}_{1,0}$'s is less than 10^{-11} .

Using the same iteration approach, $\tilde{R}_{1,i}$'s, $i > 0$ can be solved using the solution for $\tilde{R}_{1,j}$, $j < i$. The following steps are to be followed :

- i. Let initial $\tilde{R}_{1,i} = \tilde{R}_{1,i-1} \tilde{L}_1$.
- ii. Compute $\tilde{R}_{2,i}, \tilde{R}_{3,i}, \dots, \tilde{R}_{b+1,i}, \left(\tilde{R}_{j,i} = \sum_{k=0}^i \tilde{R}_k \tilde{R}_{j-1,i-k} \right)$.
- iii. Compute the new $\tilde{R}_{1,i}$ using Eqn.(E.20b).
- iv. Repeat step (ii) until maximum element-wise difference between the successive $\tilde{R}_{1,i}$'s is less than 10^{-11} .
- v. Increment i by 1 and repeat from (i).

In the study, the computation of $\tilde{R}_{1,i}$ is stopped when the iterative procedure stops in the first iteration. The desired $R_{i,j}$'s are obtained by backward transformation :

$$R_{i,j} = D'_c \tilde{R}_{i,j} (D'_c)^{-1}, \quad i \geq 0. \quad (\text{E.21})$$

Internal accuracy check is performed to verify the accuracy of the computation of R (or $\tilde{R}$). Let

$$\tilde{R}^* = \sum_{i=0}^{\infty} \tilde{R}_{1,i}, \quad (\text{E.22})$$

then from Eqn.(E.20), it follows that

$$\tilde{R}^* = \tilde{L}_2 + \tilde{R}^* \tilde{L}_1 + \sum_{i=2}^{b+1} \left(\tilde{R}^* \right)^2 \tilde{M}_{i-1}. \quad (\text{E.23})$$

Iterative method with initial $\tilde{R}^* = \text{zero matrix of size } (c+1) \times (c+1)$ is then used. The computation of the rate matrix R is considered accurate when largest element-wise difference between the $\tilde{R}^*$ obtained from Eqn.(E.22) and Eqn.(E.23) is less than 10^{-8} .

M/G/1 Paradigm and Invariant Probability Vector

Similar to Analysis 1, the invariant probability vector of process $\mathbf{Q}$, $\underline{\pi}$, is a matrix-geometric invariant vector and have the form

$$\underline{\pi} = (\underline{\pi}_{-1}, \underline{\pi}_0, \underline{\pi}_1, \underline{\pi}_2, \dots) \quad (\text{E.24})$$

which satisfies

$$\underline{\pi} \mathbf{Q} = \underline{0}, \quad \underline{\pi} \underline{e} = 1, \quad (\text{E.25})$$

and $\underline{\pi}_i$ corresponds to the probability vector on the super-block I_i , $i \geq -1$ with

$$\underline{\pi}_{i+1} = \underline{\pi}_i R, \quad i \geq 0 \quad \text{or} \quad \underline{\pi}_{i+1} = \underline{\pi}_0 R^{i+1}, \quad i \geq 0. \quad (\text{E.26})$$

The boundary probability vector $\underline{\pi}_{-1}$ and $\underline{\pi}_0$ can be obtained first by solving the following subsystem :

$$(\underline{\pi}_{-1} \quad \underline{\pi}_0) Q_{I_{-1} \cup I_0} = (\underline{0} \quad \underline{0}) \quad (\text{E.27})$$

where

$$Q_{I_{-1} \cup I_0} = \begin{bmatrix} B_{-1,-1} & B_{-1,0} \\ \sum_{i=0}^{b-1} R^i B_{i,-1} & \sum_{i=0}^{b-1} R^i B_{i,0} + R^b A_{b+1} \end{bmatrix} \quad (\text{E.28})$$

and

$$\begin{aligned}
 \pi_{-1} &= (\pi_{-1,0}, \pi_{-1,1}, \dots, \pi_{-1,c-1}) \\
 \pi_0 &= (\pi_{0,0}, \pi_{1,0}, \pi_{2,0}, \dots) \\
 \pi_{-1,i} &= \text{Prob. vector on states in block } H_i, 0 \leq i \leq c-1 \\
 \pi_{i,0} &= \text{Prob. vector on states in block } H_{i,0}, i \geq 0
 \end{aligned} \tag{E.29}$$

Note : $Q_{I_{-1} \cup I_0}$ is a matrix with negative diagonal entries, non-negative off-diagonal entries, and row sums equal to 0; hence it can be treated as a transition rate matrix for the Markov process on $I_{-1} \cup I_0$.

The matrix $\left(\sum_{i=0}^{b-1} R^i B_{i,0} + R^b A_{b+1} \right)$ has the following structure :

$$\left(\sum_{i=0}^{b-1} R^i B_{i,0} + R^b A_{b+1} \right) = \begin{bmatrix} C_1 & C_2 & C_3 & C_4 & \dots \\ C_0 & C_1 & C_2 & C_3 & \dots \\ 0 & C_0 & C_1 & C_2 & \dots \\ 0 & 0 & C_0 & C_1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{E.30}$$

with each C_i 's, $i \geq 0$ is a square matrix of order $(c+1)$ where

$$\begin{aligned}
 C_0 &= M_0 + \sum_{j=1}^{b-1} R_{j,0} Z_j, \\
 C_1 &= -D_c + \sum_{j=1}^b R_{j,0} M_j + \sum_{j=1}^{b-1} R_{j,1} Z_j, \\
 C_2 &= L_1 + \sum_{j=1}^b R_{j,1} M_j + \sum_{j=1}^{b-1} R_{j,2} Z_j, \\
 C_i &= \sum_{j=1}^b R_{j,i-1} M_j + \sum_{j=1}^{b-1} R_{j,i} Z_j, \quad i \geq 3.
 \end{aligned} \tag{E.31}$$

The matrix $\sum_{i=0}^{b-1} R^i B_{i,-1}$ has the structure

$$\sum_{i=0}^{b-1} R^i B_{i,-1} = \begin{bmatrix} 0 & 0 & 0 & \cdots & Y_b & Y_{b-1} & \cdots & Y_1 \\ 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots \end{bmatrix} \quad (\text{E.32})$$

where

$$Y_k = \begin{cases} F_b & \text{if } k = b \\ X'_k + \sum_{i=1}^{b-k} R_{i,0} X'_{i+k} & \text{if } 2 \leq k \leq b-1, k > \text{rem}(c, b) \\ X_k + \sum_{i=1}^{b-k} R_{i,0} X_{i+k} & \text{if } 2 \leq k \leq b-1, k \leq \text{rem}(c, b) \\ J_c + \sum_{i=1}^{b-1} R_{i,0} X_{i+1} & \text{if } k = 1 \end{cases} \quad (\text{E.33})$$

for $1 \leq k \leq b$.

The matrix $\left(\sum_{i=0}^{b-1} R^i B_{i,0} + R^b A_{b+1} \right)$ has M/G/1 paradigm. In order to solve the boundary vector $(\pi_{-1} \ \pi_0)$, the Markov process on $I_{-1} \cup I_0 \left(= \left(\bigcup_{m=0}^{c-1} H_m \right) \cup \left(\bigcup_{i=0}^{\infty} H_{i,0} \right) \right)$ is converted to an embedded Markov chain on $I_{-1} \cup I_0$ and the boundary vector (with finite size) $(\pi_{-1}, \pi_{0,0})$ can first be determined. The transformation from the Markov process to the embedded Markov chain on $I_{-1} \cup I_0$ can be done by letting :

$$\begin{aligned} \tilde{C}_1 &= D_c^{-1} C_1 + I, \ I \text{ is identity matrix of order } (c+1) \\ \tilde{C}_i &= D_c^{-1} C_i, \ i = 0, 2, 3, 4, \dots \\ \tilde{K}_i &= D_i^{-1} K_i, \ 0 \leq i \leq c \\ \tilde{J}_i &= D_i^{-1} J_i, \ 0 \leq i \leq c \\ \tilde{F}_i &= D_i^{-1} F_i, \ 0 \leq i \leq c \\ \tilde{E}_i &= D_i^{-1} E_i, \ 0 \leq i \leq c \\ \tilde{Y}_i &= D_i^{-1} Y_i, \ 1 \leq i \leq b. \end{aligned} \quad (\text{E.34})$$

Following Lucantoni's [2] and Miller's [20] approach, the following matrices are determined first :

1. G : a square matrix of order $(c+1)$ denoting the hitting probabilities of states in $H_{j,c}$ starting at $H_{j+1,0}$, $j \geq 0$ with G satisfies

$$G = \sum_{v=0}^{\infty} \tilde{C}_v G^v. \quad (\text{E.35})$$

G can be found iteratively following Step 3 in Appendix A.

2. V : a square matrix of order $(c+1)$ denoting the first passage probabilities of hitting states in $H_{j+1,0}$ starting from states in $H_{j,0}$ without first return to $H_{j,0}$, $j \geq 0$. V satisfies

$$V = \sum_{v=0}^{\infty} \tilde{C}_v G^{v-2}. \quad (\text{E.36})$$

3. W : a square matrix of order $(c+1)$ denoting the first passage probabilities of hitting states in $H_{j+1,0}$ starting from states in $H_{j+1,0}$ without visiting to $H_{j,0}$, $j \geq 0$. W satisfies

$$W = \sum_{v=1}^{\infty} \tilde{C}_v G^{v-1}. \quad (\text{E.37})$$

4. $\tilde{S}_i$, $i \geq 1$: a square matrix of order $(c+1) \times b$ whose entries are the expected number of visits to states in $H_{j+1,0}$ starting from states in $H_{j,0}$ before hitting the states in $H_{k,0}$, $k < j + i$. $\tilde{S}_i$, $i \geq 1$ is computed as

$$\tilde{S}_i = \sum_{v=i+1}^{\infty} \tilde{C}_v G^{v-i+1} (I - W)^{-1} \quad (\text{E.38})$$

where I is an identity matrix of the same order.

then they are obtained by denormalization shown as follows :

$$\pi_{-1,i} = \left(\frac{1-\rho}{\xi} \right) \underline{y}_{-1,i}, \quad 0 \leq i \leq c-1, \quad \text{and} \quad \pi_{0,0} = \left(\frac{1-\rho}{\xi} \right) \underline{y}_{0,0} \quad (\text{E.43})$$

where

$$\begin{aligned} \rho &= \text{system utilization} = \frac{b\lambda_1}{c\mu_1} + \frac{\lambda_2}{c\mu_2}, \quad \text{and} \\ \xi &= \sum_{i=0}^{c-1} \frac{c-i}{c} \underline{y}_{-1,i} \underline{e}, \\ \underline{e} &= \text{unit column vector of size } \left\lfloor \frac{i}{b} \right\rfloor. \end{aligned} \quad (\text{E.44})$$

Then the rest of the probability vectors in super-state I_0 , $\pi_{1,0}$, $\pi_{2,0}$, $\pi_{3,0}$, ..., $\pi_{\infty,0}$, can be obtained as

$$\pi_{i+1,0} = \sum_{j=0}^i \pi_{j,0} D_c \bar{S}_{i+1-j} D_c^{-1}, \quad i \geq 0. \quad (\text{E.45})$$

For $\pi_{i,j}$, $i, j \geq 0$, the entries of the rate matrix R are employed for the computation using the following equations :

$$\pi_{i,j+1} = \sum_{k=0}^i \pi_{k,j} R_{i,k-j}. \quad (\text{E.46})$$

The computation of the invariant probability vector can be truncated to any level. In the implementation, the truncation level is determined when $\sum_{i=0}^{\infty} \pi_{i,j+1} \underline{e} < 10^{-9}$ is satisfied.

Appendix F

Algorithmic Solutions For Preemptive Priority M/M/c Model (Analysis 3 — Sect. 4.5)

Followings are the procedures for the invariant probability vector for the M/M/c queueing model where Type 1-SRs have a preemptive priority over Type 2-SR in c_1 ($\leq c$) servers only and that Type 2-SRs have a preemptive priority over Type 1-SR in the remaining c_2 ($= c - c_1$) servers.

Define n_i ($i = 1, 2$) is the number of Type i -SRs in the system, then the state space of the system is given as

$$S = \{(n_1, n_2), n_1, n_2 \geq 0\}, \quad (\text{F.1})$$

and define

$$p_{i,j} = \text{Prob. } \{n_1 = i, n_2 = j\}. \quad (\text{F.2})$$

By combining the set of states $\{(n_1 = i, n_2 = j) \mid j \geq 0\}$ to form a super-state, H_i ($i \geq 0$), the corresponding probability vector for that super-state is

$$\underline{p}_i = (p_{i,0}, p_{i,1}, p_{i,2}, \dots, p_{i,\infty}). \quad (\text{F.3})$$

The system can be modelled as a Markov process with the transition rate matrix Q given by

$$Q = \begin{bmatrix} B_0 & D_0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ C_1 & B_1 & D_1 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & C_2 & B_2 & D_2 & 0 & 0 & 0 & 0 & \dots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & C_r & B_r & D_r & 0 & 0 & \dots \\ 0 & 0 & \dots & 0 & A_2 & A_1 & A_0 & 0 & \dots \\ 0 & 0 & \dots & 0 & 0 & A_2 & A_1 & A_0 & \dots \\ \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \quad (F.4)$$

and its invariant probability vector $\underline{p}$ given by

$$\underline{p} = (\underline{p}_0, \underline{p}_1, \underline{p}_2, \dots) \quad (F.5)$$

which satisfies

$$\underline{p}Q = \underline{0} \text{ and } \underline{p}\underline{e} = 1. \quad (F.6)$$

Note : $r = \left\lceil \frac{c}{b} \right\rceil - 1$ where $\lceil \cdot \rceil$ is the smallest integer not less than the argument, and $\underline{0}$ is an infinite size zero vector and $\underline{e}$ is an infinite size unit vector.

Each of the submatrices A_i 's, B_i 's, C_i 's is 2 dimensional matrix of infinite size and the

elements are given as follows :

$A_0 = \lambda_1 I$, where I is an identity matrix

$$\begin{aligned}
 A_1(i, j) &= \begin{cases} \lambda_2 & \text{if } j = i + 1 \\ -(\lambda + \zeta\mu_1 + \min(j, c - b\zeta)\mu_2) & \text{if } j = i \\ \min(j, c - b\zeta)\mu_2 & \text{if } j = i - 1 \\ 0 & \text{otherwise} \end{cases} \\
 &\text{for } 0 \leq i, j \leq \infty \\
 A_2(i, j) &= \begin{cases} \zeta\mu_1 & \text{if } j = i \\ 0 & \text{otherwise} \end{cases} \\
 &\text{for } 0 \leq i, j \leq \infty \\
 B_m(i, j) &= \begin{cases} \lambda_2 & \text{if } j = i + 1 \\ -(\lambda + \min(\zeta, m)\mu_1 + \min(j, c - b(\min(\zeta, m)))\mu_2) & \text{if } j = i \\ \min(i, c - b(\min(\zeta, m)))\mu_2 & \text{if } j = i - 1 \\ 0 & \text{otherwise} \end{cases} \\
 &\text{for } 0 \leq m \leq r, 0 \leq i, j \leq \infty \\
 C_m(i, j) &= \begin{cases} \min(\zeta, m)\mu_1 & \text{if } j = i \\ 0 & \text{otherwise} \end{cases} \\
 &\text{for } 1 \leq m \leq r, 0 \leq i, j \leq \infty \\
 D_m &= A_0, 0 \leq m \leq r
 \end{aligned} \tag{F.7}$$

where $\zeta = \frac{c - \min(j, c_2)}{b}$ and $\lambda = \lambda_1 + \lambda_2$.

The Markov process shown is the QBD process. From Neuts' method [1], the stationary probability vector for this system (which has a stable queue), $\underline{p}$, partitioned into probability vectors which correspond to each super-states, is given as

$$\underline{p} = (\underline{p}_0, \underline{p}_1, \underline{p}_2, \dots, \underline{p}_{r-1}, \underline{p}_r, \underline{p}_{r+1}, \dots) \tag{F.8}$$

where

$$\underline{p}_i = \underline{p}_r R^{i-r}, \text{ for } i \geq r \tag{F.9}$$

and the rate matrix R is the minimal non-negative solution to the equation

$$R^2 A_2 + R A_1 + A_0 = 0. \quad (\text{F.10})$$

Determination of R

Since the system is designed such that Type 1-SR and Type 2-SR can preempt each other in their designated regions, it is possible for the system to start in state $(n_1 = i, n_2 = v), v > 0$, and then on the first return to the set $\{(n_1 = i, n_2 = k), k \geq 0\}$ after visiting state $(n_1 = i + 1, n_2 = j), j \geq 0$, it reaches the state $(n_1 = i, n_2 = w), w < v$. The nice upper-triangular structure of R , which appears in Analysis 1 and Analysis 2, therefore does not exist and the evaluation of infinite size R cannot be performed by partitioning the matrix and evaluate each small partitioned block separately. For numerical analysis to be possible, the infinite size of the aforementioned matrices, i.e. A_i, B_i, C_i etc., have to be arbitrary fixed at a large enough size such that accuracy of the algorithm is acceptable while maintaining the efficiency of the algorithm. Since the infinite size of the above matrices corresponds to the infinite system size of Type 2-SR, the crude approximation for the finite minimum Type 2-SR system length required is described below :

1. Case with $\lambda_2/c_2\mu_2 < 1$: in the region of c_2 servers, the Type 2-SR has absolute preemptive priority over Type 1-SR, hence by solving a simple M/M/ c_2 system with λ_2 and μ_2 , the minimum system length required can be determined. In the implementation, the system length is fixed at i where $\text{prob.}(i \text{ Type 2-SR in M/M}/c_2 \text{ system}) < 10^{-12}$.
2. Case with $\lambda_2/c_2\mu_2 \geq 1$: the number of servers that is available for Type 2-SR is approximately $c - \left\lfloor \frac{b\lambda_1}{\mu_1} \right\rfloor$. By solving the simple M/M/ $(c - \left\lfloor \frac{b\lambda_1}{\mu_1} \right\rfloor)$ system with λ_2 and μ_2 , the minimum system length required is chosen as $3 \times i$ where $\text{prob.}(i \text{ Type 2-SR}$

in $M/M/(c - \lfloor \frac{b\lambda_1}{\mu_1} \rfloor)$ system) $< 10^{-12}$. Although this approximation is very crude but it provides fairly accurate approximation on the minimum system length required.

After the size of R , as well as other matrices, has determined, the tranformation from Markov process to embedded Markov chain is required in order to evaluate R . By letting

$$\begin{aligned}\tilde{A}_0 &= \Lambda^{-1} A_0, \\ \tilde{A}_1 &= \Lambda^{-1} A_1 + I, \\ \tilde{A}_2 &= \Lambda^{-1} A_2, \text{ and} \\ \tilde{R} &= \Lambda^{-1} R \Lambda,\end{aligned}\tag{F.11}$$

where Λ is a diagonal matrix with element $(\Lambda)_{jj} = -(A_1)_{jj}$, Eqn.(F.10) becomes

$$\tilde{R}^2 \tilde{A}_2 + \tilde{R} \tilde{A}_1 + \tilde{A}_0 = \tilde{R}.\tag{F.12}$$

Modified Newton-Kantorovich iteration scheme is used to evaluate $\tilde{R}$ (see Appendix A : Step 3). R can be obtained by backward transformation as $R = \Lambda R \Lambda^{-1}$.

Boundary probability vectors

The boundary probability vectors $\underline{p}_0, \underline{p}_1, \underline{p}_2, \dots, \underline{p}_{r-1}, \underline{p}_r$ satisfy either one of the following sets of equations :

$$\begin{aligned}\text{for } r > 0 : & \begin{cases} \underline{p}_0 B_0 + \underline{p}_1 C_1 = \underline{0} \\ \underline{p}_{i-1} D_{i-1} + \underline{p}_i B_i + \underline{p}_{i+1} C_{i+1} = \underline{0}, & \text{for } 1 \leq i \leq r-1 \\ \underline{p}_{r-1} D_{r-1} + \underline{p}_r (B_r + R A_2) = \underline{0} \\ \sum_{v=0}^{r-1} \underline{p}_v \underline{e} + \underline{p}_r (I - R)^{-1} \underline{e} = 1 \end{cases} \\ \text{for } r = 0 : & \begin{cases} \underline{p}_0 B_0 + \underline{p}_1 A_1 = \underline{0} \\ \underline{p}_0 A_0 + \underline{p}_1 (A_1 + R A_2) = \underline{0} \\ \underline{p}_0 (I - R)^{-1} \underline{e} = 1 \end{cases}\end{aligned}\tag{F.13}$$

For computational efficiency, the probability vectors $\underline{p}_i$ ($0 \leq i \leq r$) are also obtained by solving the corresponding probability vectors of the embedded Markov chain and appropriately transforming those of the embedded Markov chain back. The corresponding set of equations for the embedded Markov chain is

$$\begin{aligned} \text{for } r > 0 : \quad & \begin{cases} \tilde{\underline{p}}_0 \tilde{B}_0 + \tilde{\underline{p}}_1 \tilde{C}_1 = \tilde{\underline{p}}_0 \\ \tilde{\underline{p}}_{i-1} \tilde{D}_{i-1} + \tilde{\underline{p}}_i \tilde{B}_i + \tilde{\underline{p}}_{i+1} \tilde{C}_{i+1} = \tilde{\underline{p}}_i, & \text{for } 1 \leq i \leq r-1 \\ \tilde{\underline{p}}_{r-1} \tilde{D}_{r-1} + \tilde{\underline{p}}_r (\tilde{B}_r + \tilde{R} \tilde{A}_2) = \tilde{\underline{p}}_r \\ \sum_{v=0}^{r-1} \tilde{\underline{p}}_v \underline{e} + \tilde{\underline{p}}_r (I - \tilde{R})^{-1} \underline{e} = 1 \end{cases} \\ \text{for } r = 0 : \quad & \begin{cases} \tilde{\underline{p}}_0 \tilde{B}_0 + \tilde{\underline{p}}_1 \tilde{A}_1 = \tilde{\underline{p}}_0 \\ \tilde{\underline{p}}_0 \tilde{A}_0 + \tilde{\underline{p}}_1 (\tilde{A}_1 + \tilde{R} \tilde{A}_2) = \tilde{\underline{p}}_1 \\ \tilde{\underline{p}}_0 (I - \tilde{R})^{-1} \underline{e} = 1 \end{cases} \end{aligned} \quad (\text{F.14})$$

where

$$\begin{aligned} \tilde{B}_k &= \Lambda_k^{-1} B_k + I, \\ \tilde{C}_k &= \Lambda_k^{-1} C_k, \\ \tilde{D}_k &= \Lambda_k^{-1} D_k, \quad 0 \leq k \leq r \end{aligned} \quad (\text{F.15})$$

and Λ_k , $0 \leq k \leq r$ is the diagonal matrix with the elements $(\Lambda_k)_{jj} = (B_k)_{jj}$.

The set of equations is solved by iteration assuming the initial set of vectors $(\tilde{\underline{p}}_0, \tilde{\underline{p}}_1, \tilde{\underline{p}}_2, \dots, \tilde{\underline{p}}_{r-1}, \tilde{\underline{p}}_r)$ and with row sum equals one. Iteration stops until maximum element-wise difference in $(\tilde{\underline{p}}_0, \tilde{\underline{p}}_1, \tilde{\underline{p}}_2, \dots, \tilde{\underline{p}}_{r-1}, \tilde{\underline{p}}_r)$ between successive iterations is less than an epsilon 10^{-11} . The actual boundary vectors $\underline{p}_0, \underline{p}_1, \underline{p}_2, \dots, \underline{p}_{r-1}, \underline{p}_r$ are then obtained as follows :

$$\underline{p}_i = \tilde{\underline{p}}_i D_i^{-1} \text{ where } 0 \leq i \leq r. \quad (\text{F.16})$$

The rest of the probability vectors can be obtained as

$$\underline{p}_i = \underline{p}_r R^{i-r} \text{ for } i \geq r. \quad (\text{F.17})$$

Since the system is a preemptive priority system, the number of Type i -SRs ($i = 1, 2$) *waiting* in the respective infinite size queues are readily known from the number of Type i -SRs in the system. Denote s_i , $i = 1, 2$ as the number of servers Type i -SR is occupying and w_i , $i = 1, 2$ as the number of Type i -SR waiting in the respective queue. Given

$$n_1 = i \text{ and } n_2 = j, \quad i, j \geq 0 \quad (\text{F.18})$$

then

$$\begin{aligned} s_1 &= \min(c - \min(j, c_2), \min(ib, c)), \\ s_2 &= \min(j, \max(c - ib, c_2)) \text{ and} \\ \text{No. of idle servers} &= c - s_1 - s_2. \end{aligned} \quad (\text{F.19})$$

If there is idle server, obviously $w_1 = w_2 = 0$; otherwise, $w_1 = i - \left\lceil \frac{s_1}{b} \right\rceil$ and $w_2 = j - s_2$.

Appendix G

Procedures for Approximating Videophone Sources given as Minisources with MMPP

The procedures given in this appendix is summarized from [23]. Readers should refer to [23] for details.

The following statistical characteristics of the superposition of n minisources are matched with those of MMPP to yield the four parameters (λ_1 , λ_2 , r_1 and r_2) of the 2-state MMPP :

1. mean packet arrival rate
2. variance-to-mean ratio of the number of packet arrival in $(0, t_1)$
3. long term variance-to-mean ratio of the number of packet arrivals, and
4. third central moment of the number of packet arrivals in $(0, t_2)$

Minisources

Each minisource derived from the discrete-state continuous-time Markov process with parameters α , β , and A evaluated from Eqn.(5.1) and Eqn.(5.2) is considered as a renewal process. Renewal theory results are used to evaluate the mean, variance-

mean ratio, and third central moment of the number of arrivals in a time interval for a superposition of n packetized videophone minisources [23].

The packet stream from one minisource is modelled by arrivals at fixed intervals of T sec. ($T=1/A$ where A is coder output rate in packet/sec.) during ON state and no arrivals during OFF state (see Fig. 5.4). The renewal process (packet arrival process) has interarrival time distribution given by

$$F(t) = \left[(1 - \alpha T) + \alpha T (1 - e^{-\beta(t-T)}) \right] U(t - T) \quad (\text{G.1})$$

where $U(t)$ is the unit step function and the Laplace Stieltjes transform (LST) is given by

$$\tilde{f}(s) = \int_0^{\infty} e^{-st} dF(t) = [1 - \alpha T + \alpha T \beta / (s + \beta)] e^{-sT} \quad (\text{G.2})$$

with the mean packet arrival rate from a single source given by

$$\lambda = -1/\tilde{f}'(0) = 1/(T + \alpha T/\beta) = A\beta/(\alpha + \beta) \quad (\text{G.3})$$

Let $N(0, t)$ denotes the number of arrivals of a stationary renewal process in the interval $(0, t)$ and define $M_r(t) = E[N^r(0, t)]$, and $\tilde{M}_r(s) = \mathcal{L}[M_r(t)]$ where $\mathcal{L}(\cdot)$ denotes the Laplace transform. Then

$$\begin{aligned} \tilde{M}_1(s) &= \frac{\lambda}{s^2} \\ \tilde{M}_2(s) &= \frac{\lambda}{s^2} \left(\frac{1 + \tilde{f}(s)}{1 - \tilde{f}(s)} \right) \\ \tilde{M}_3(s) &= \frac{\lambda}{s^2} \left(\frac{1 + 4\tilde{f}(s) + \tilde{f}^2(s)}{(1 - \tilde{f}(s))^2} \right) \end{aligned} \quad (\text{G.4})$$

where $\tilde{f}(s)$ is the LST of the interarrival time distribution and λ^{-1} is the mean time between arrivals.

As shown in [23],

$$M_1(t) = t/[T + \alpha T/\beta] = \lambda t \quad \text{and} \quad \lim_{t \rightarrow \infty} \frac{\text{var}(N(0, t))}{M_1(t)} = \frac{1 - (1 - \alpha T)^2}{(\alpha T + \beta T)^2} \quad (\text{G.5})$$

The second and third moments of the number of arrivals in a *finite-time* interval is obtained numerically via numerical Laplace transform inversion [29] of $\tilde{M}_2(s)$, $\tilde{M}_3(s)$ at the desired time t .

Let the number of packet arrivals in $(0, t)$ for the superposition of n minisources be

$$N^S(0, t) = \sum_{i=1}^n N_i(0, t) \quad (\text{G.6})$$

where $N_i(0, t)$ is the number of packet arrivals in $(0, t)$ from the i^{th} minisource. Hence

$$M_1^S(t) = E[N^S(0, t)] = nM_1(t) \quad (\text{G.7})$$

and

$$\frac{\text{var}[N^S(0, t)]}{E[N^S(0, t)]} = \frac{\text{var}[N(0, t)]}{E[N(0, t)]} = \frac{M_2(t) - (M_1(t))^2}{M_1(t)} \quad (\text{G.8})$$

where superscript S denotes superposition variables.

The third central moments for the superposition process is

$$\begin{aligned} \mu_3^{*S}(0, t) &= E\left\{ \left[N^S(0, t) - E(N^S(0, t)) \right]^3 \right\} \\ &= n[M_3(t) - 3M_2(t)M_1(t) + 2M_1^3(t)] \end{aligned} \quad (\text{G.9})$$

where $M_2(t)$ and $M_3(t)$ are obtained from Laplace transform inversion of $\tilde{M}_2(s)$ and $\tilde{M}_3(s)$ at desired time.

Following the steps above, the quantities of the aggregation of n minisources required for matching the MMPP are available, namely

$$\begin{aligned} M_1^S(t) &= n\lambda t, \\ \text{var}(N^S(0, t_1))/E[N^S(0, t_1)], \\ \lim_{t \rightarrow \infty} \left(\text{var}(N^S(0, t_1))/E[N^S(0, t_1)] \right), \text{ and} \\ \mu_3^{*S}(0, t_2). \end{aligned} \quad (\text{G.10})$$

Markov Modulated Poisson Process (MMPP)

The expressions for the four quantities (statistical characteristics) for the 2-state MMPP are obtained using the probability generating function of the number of arrivals in an interval. Detailed derivation is given in [23]. The probability generating function is

$$g(z, t) = \pi \exp \{ [Q^* + (z - 1)\Lambda]t \} \underline{e}, \quad (\text{G.11})$$

where

$$\pi = \frac{1}{r_1 + r_2} (r_2, r_1), \quad \underline{e} = (1, 1)^T$$

$$Q^* = \begin{bmatrix} -r_1 & r_1 \\ r_2 & -r_2 \end{bmatrix}, \quad \Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \quad (\text{G.12})$$

Let N_t denotes the number of packet arrivals of the stationary 2-state MMPP over the interval $(0, t)$, then

$$\overline{N_t} = E[N_t] = \frac{\lambda_1 r_2 + \lambda_2 r_1}{r_1 + r_2} t, \quad (\text{G.13})$$

and

$$\frac{\text{var}(N_t)}{E[N_t]} = 1 + \frac{2(\lambda_1 - \lambda_2)^2 r_1 r_2}{(r_1 + r_2)^2 (\lambda_1 r_2 + \lambda_2 r_1)} - \frac{2(\lambda_1 - \lambda_2)^2 r_1 r_2}{(r_1 + r_2)^3 (\lambda_1 r_2 + \lambda_2 r_1) t} \cdot (1 - e^{-(r_1 + r_2)t}), \quad (\text{G.14})$$

and

$$\lim_{t \rightarrow \infty} \frac{\text{var}(N_t)}{E[N_t]} = 1 + \frac{2(\lambda_1 - \lambda_2)^2 r_1 r_2}{(r_1 + r_2)^2 (\lambda_1 r_2 + \lambda_2 r_1)}, \quad (\text{G.15})$$

and

$$E[(N_t - \overline{N_t})^3] = g^{(3)}(1, t) - 3\overline{N_t}(\overline{N_t} - 1) \frac{\text{var}(N_t)}{\overline{N_t}} - \overline{N_t}(\overline{N_t} - 1)(\overline{N_t} - 2) \quad (\text{G.16})$$

where

$$g^{(3)}(1, t) = \frac{6}{r_1 + r_2} \left[\frac{A_{11}}{6} t^3 + \frac{A_{21}}{2} t^2 + A_{31} t + A_{12} t e^{-(r_1 + r_2)t} + A_{41} (1 - e^{-(r_1 + r_2)t}) \right], \quad (\text{G.17})$$

and

$$\begin{aligned}
 A_{11} &= \frac{(\lambda_1 r_2 + \lambda_2 r_1)^3}{(r_1 + r_2)^2}, \\
 A_{21} &= \frac{2r_1 r_2 (\lambda_1 - \lambda_2)^2 (\lambda_1 r_2 + \lambda_2 r_1)}{(r_1 + r_2)^3}, \\
 A_{31} &= \frac{r_1 r_2 (\lambda_1 - \lambda_2)^2 [\lambda_1 r_1 + \lambda_2 r_2 - 2(\lambda_1 r_2 + \lambda_2 r_1)]}{(r_1 + r_2)^4}, \\
 A_{41} &= \frac{-2r_1 r_2 (\lambda_1 - \lambda_2)^3 (r_1 - r_2)}{(r_1 + r_2)^5}, \quad \text{and} \\
 A_{12} &= \frac{r_1 r_2 (\lambda_1 - \lambda_2)^2 (\lambda_1 r_1 + \lambda_2 r_2)}{(r_1 + r_2)^4}.
 \end{aligned} \tag{G.18}$$

Determination of r_1 , r_2 , λ_1 and λ_2 of the 2-state MMPP

By matching the statistical characteristics, the following steps lead to the determination of the four parameters of the 2-state MMPP at arbitrary time t_1 and t_2 .

1. Denote

$$\begin{aligned}
 a &= \frac{E[N^S(0, t)]}{t}, \\
 b_t &= \frac{\text{var}[N^S(0, t)]}{E[N^S(0, t)]} = \frac{\text{var}[N(0, t)]}{E[N(0, t)]}, \quad \text{and} \\
 b_\infty &= \lim_{t \rightarrow \infty} \frac{\text{var}[N^S(0, t)]}{E[N^S(0, t)]}
 \end{aligned} \tag{G.19}$$

all the above quantities can be computed since they are known from the aggregation of minisources.

2. Obtain

$$g^{(3)}(1, t_2) = \mu^{*S}(0, t_2) + 3at_2(at_2 - 1)b_{t_2} + at_2(at_2 - 1)(at_2 - 2) \tag{G.20}$$

at a specific time t_2 .

3. Compute b_t at $t = t_1$, then if $b_{t_1} > 1$, i.e. the variance-to-mean ratio of the superposition process at t_1 is greater than that of a Poisson process, then $(r_1 +$

r_2) can be solved by letting $d = (r_1 + r_2)$ and successive iteration of

$$d = \frac{1}{t_2} \left(\frac{b_\infty - 1}{b_\infty - b_{t_1}} \right) (1 - e^{-dt_1}). \quad (\text{G.21})$$

4. Denotes $K = (\lambda_1 - \lambda_2)(r_1 - r_2)$, then K can be obtained by rearranging the equation given by :

$$\begin{aligned} g^{(3)}(1, t_2) = & a^3 t_2^3 + 3a^2(b_\infty - 1)t_2^2 + \frac{3a(b_\infty - 1)}{d} \left(\frac{K}{d} - a \right) t_2 \\ & + \frac{3a}{d^2} (b_\infty - 1)(K + ad)t_2 e^{-dt_2} \\ & - \frac{6a}{d^3} (b_\infty - 1)K(1 - e^{-dt_2}) \end{aligned} \quad (\text{G.22})$$

5. If $K = 0$, then $\lambda_1 = \lambda_2 = d/2$, otherwise compute $e = \frac{(b_\infty - 1)ad^3}{2K^2}$.

6. The parameters of the 2-state MMPP is given by :

$$\begin{aligned} r_1 &= \frac{d}{2} \left(1 + \frac{1}{\sqrt{4e + 1}} \right) \\ r_2 &= d - r_1 \\ \lambda_2 &= \left(\frac{ad}{r_2} - \frac{K}{r_1 - r_2} \right) \left(\frac{r_2}{r_1 + r_2} \right) \\ \lambda_1 &= \frac{K}{r_1 - r_2} + \lambda_2. \end{aligned} \quad (\text{G.23})$$

Note : The time point t_1 and t_2 should be chosen such that the variance-to-mean ratio of the 2-state MMPP is a good fit to that of the superposition of minisources.

Appendix H

Matrix Analytic Solution for MMPP/D/c Queueing System

The following is the algorithm used to solve the invariant probability vector for a MMPP/D/c Queueing system found in Chapter 5 — Sect. 5.3.1. This algorithm is reported in [25] by M.F. Neuts and is summarized here. The arrival process is a m -state MMPP. The basic approach is to partition the transition probability matrix of the Markov chain into a Markov chain of M/G/1 Type and apply the known solution.

The transition probability matrix $\mathbf{P}$ (as in Eqn.(5.16)) has the form

$$\mathbf{P} = \begin{matrix} & \begin{matrix} \langle 0 \rangle \\ \langle 1 \rangle \\ \vdots \\ \langle c-1 \rangle \\ \langle c \rangle \\ \langle c+1 \rangle \\ \langle c+2 \rangle \\ \langle c+3 \rangle \\ \vdots \end{matrix} & \begin{bmatrix} P(0,T) & P(1,T) & P(2,T) & P(3,T) & \cdots \\ P(0,T) & P(1,T) & P(2,T) & P(3,T) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ P(0,T) & P(1,T) & P(2,T) & P(3,T) & \cdots \\ P(0,T) & P(1,T) & P(2,T) & P(3,T) & \cdots \\ 0 & P(0,T) & P(1,T) & P(2,T) & \cdots \\ 0 & 0 & P(0,T) & P(1,T) & \cdots \\ 0 & 0 & 0 & P(0,T) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \end{matrix} \quad (\text{H.1})$$

$\mathbf{P}$ is further partitioned into

$$\mathbf{P} = \begin{matrix} & \begin{matrix} \underline{0} \\ \underline{1} \\ \underline{2} \\ \underline{3} \\ \vdots \end{matrix} & \begin{bmatrix} B_0 & B_1 & B_2 & B_3 & \cdots \\ A_0 & A_1 & A_2 & A_3 & \cdots \\ 0 & A_0 & A_1 & A_2 & \cdots \\ 0 & 0 & A_0 & A_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \end{matrix} \quad (\text{H.2})$$

where B_i 's, A_i 's, $i \geq 0$ are given by Eqn.(5.18).

The joint steady state probability is defined as

$$x_{k,l} = \Pr[{}_H Q_n = k, H_n = l], \quad k \geq 0, \quad 1 \leq l \leq m \quad (\text{H.3})$$

where the definition of ${}_H Q_n$ and H_n is given in Eqn.(5.8). The invariant probability vector is then defined as

$$\underline{x} = (\tilde{x}_0, \tilde{x}_1, \tilde{x}_2, \dots) \quad (\text{H.4})$$

where

$$\begin{aligned} \tilde{x}_k &= (x_{kc}, x_{kc+1}, x_{kc+2}, \dots, x_{(k+1)c-1}), \quad k \geq 0 \text{ and} \\ x_{kc+i} &= (x_{kc+i,1}, x_{kc+i,2}, \dots, x_{kc+i,m}), \quad 0 \leq i \leq c-1 \end{aligned} \quad (\text{H.5})$$

and satisfies the followings :

$$\underline{x} \mathbf{P} = \underline{x}, \quad \underline{x} \underline{e} = 1. \quad (\text{H.6})$$

Procedures for solving the invariant vector :

1. Solve for $P(v, T)$, $v \geq 0$ via numerical integration of the following set of differential equations :

$$\begin{aligned} P'(0, T) &= P(0, T)[Q^* - \Lambda] \\ P'(v, T) &= P(v, T)[Q^* - \Lambda] + P(v-1, T)\Lambda \quad \text{for } v \geq 1 \end{aligned} \quad (\text{H.7})$$

with initial conditions $P(v, 0) = \delta_{v0} I$, for $v \geq 0$. Runge-Kutta method is used for numerical integration [27] and matrices Q^* and Λ are given by Eqn.(5.4)

and Eqn.(5.5). They can also be generalized for m -state MMPP (see [24, 25]). Truncation for $P(v, T)$, $v \geq 0$ is determined when the largest element is less than 10^{-12} .

2. Determine the invariant probability vector $\underline{\pi}^*$ by iteration where

$$\underline{\pi}^* A = \underline{\pi}^*, \quad \underline{\pi}^* \underline{e} = c, \quad A = \sum_{v=0}^{\infty} A_v \quad (\text{H.8})$$

and $\underline{e} = (1, 1, 1, \dots, 1)^T$ is a column vector of an appropriate dimension.

Note that $\underline{\pi}^* = c^{-1}[\underline{\pi}, \underline{\pi}, \dots, \underline{\pi}]$ and $\underline{\pi}$ is the invariant probability of Q^* which satisfies

$$\underline{\pi} Q^* = \underline{0}, \quad \underline{\pi} \underline{e} = 1. \quad (\text{H.9})$$

* Power method [27] is suggested in the evaluation of the invariant vector.

3. Find the vector $\underline{\beta}^*$: $\underline{\beta}^* = \sum_{k=1}^{\infty} k A_k \underline{e}.$

4. Check if

$$\underline{\pi}^* \underline{\beta}^* = \rho (< 1) = c^{-1} \mu^* T \quad (\text{H.10})$$

where ρ is the system utilization and $\mu^* = \underline{\pi} \Lambda \underline{e}$ (stationary packet arrival rate of the MMPP).

5. Solve for $\mathbf{G}$ by iteration

$$\mathbf{G} = \sum_{k=0}^{\infty} A_k \mathbf{G}^k \quad (\text{H.11})$$

where $\mathbf{G}$ is a non-negative stochastic matrix. Detailed algorithm should be referred to Appendix A — Step 3.

It should be noted that $\mathbf{G}$ has the structure shown as follows :

$$\mathbf{G} = \begin{bmatrix} G(0,0) & G(0,1) & \cdots & G(0,c-1) \\ G(1,0) & G(1,1) & \cdots & G(1,c-1) \\ \vdots & \vdots & \ddots & \vdots \\ G(c-1,0) & G(c-1,1) & \cdots & G(c-1,c-1) \end{bmatrix} \quad (\text{H.12})$$

6. Find the invariant probability vector $\underline{g}$, where

$$\underline{g} \mathbf{G} = \underline{g}, \quad \underline{g} \underline{e} = 1. \quad (\text{H.13})$$

7. Obtain $\tilde{\mathbf{G}}$: $\tilde{\mathbf{G}} = \underline{e} \underline{g}$.

8. Find vector $\underline{\phi}$:

$$\underline{\phi} = \left(I - \mathbf{G} + \tilde{\mathbf{G}} \right) \left[I - A + \tilde{\mathbf{G}} - \Delta(\underline{\beta}^*) \tilde{\mathbf{G}} \right]^{-1} \underline{e}, \quad (\text{H.14})$$

where

$$\begin{aligned} \Delta(\underline{\beta}^*) &= \text{diag}(\beta_0, \beta_1, \beta_2, \dots, \beta_{c-1}), \text{ and} \\ \beta_i &= (\beta_{i,1}, \beta_{i,2}, \dots, \beta_{i,m}), \quad 0 \leq i \leq c-1. \end{aligned} \quad (\text{H.15})$$

I = Identity matrix of an appropriate dimension. ($\text{diag}(a, b, c, \dots)$ is diagonal matrix with corresponding elements $a, b, c, \dots$ etc.)

9. Check if

$$\underline{g} \underline{\phi} = (1 - \rho)^{-1} = c(c - \mu^* T)^{-1} \quad (\text{H.16})$$

with

$$\underline{\phi} = [\underline{\phi}_0, \underline{\phi}_1, \dots, \underline{\phi}_{c-1}]^T \quad (\text{H.17})$$

and $\underline{\phi}_i$, $0 \leq i \leq c-1$ is a row vector of size m .

10. Compute $\hat{\mathbf{G}} = \sum_{v=0}^{c-1} G(0, v)$ and its invariant vector $\hat{\underline{g}}$ which satisfies $\hat{\underline{g}} \hat{\mathbf{G}} = \hat{\underline{g}}$ and $\hat{\underline{g}} \underline{e} = 1$.

11. Compute the following matrices :

$$\begin{aligned}
 \mathbf{A}_1 &= \sum_{v=1}^{\infty} A_v \mathbf{G}^{v-1}, \\
 \mathbf{A}_2 &= \sum_{v=2}^{\infty} (v-1) A_v \tilde{\mathbf{G}} = [\underline{\beta}^* + A_0 \underline{e} - \underline{e}] \underline{g}, \\
 \mathbf{A}_3 &= \sum_{v=1}^{\infty} A_v = A - A_0, \\
 \mathbf{B}_1 &= \sum_{v=1}^{\infty} B_v \mathbf{G}^{v-1}, \\
 \mathbf{B}_2 &= \sum_{v=2}^{\infty} (v-1) B_v \tilde{\mathbf{G}}, \\
 \mathbf{B}_3 &= \sum_{v=1}^{\infty} B_v.
 \end{aligned} \tag{H.18}$$

Note that $\mathbf{B}_1$, $\mathbf{B}_2$, and $\mathbf{B}_3$ can be obtained easily by taking the first row of $m \times m$ blocks in the corresponding matrix $\mathbf{A}_i$, $i = 1, 2, 3$ and repeat the row c times.

12. Compute the stochastic matrix K :

$$K = A_0(I - B_0)^{-1} \mathbf{B}_1 + \mathbf{A}_1 \tag{H.19}$$

and its invariant vector $\underline{k}$ where

$$\underline{k} = \underline{k} K, \quad \underline{k} \underline{e} = 1. \tag{H.20}$$

13. Compute the inverse matrix of $(I - B_0)$:

$$(I - B_0)^{-1} = \begin{bmatrix} I + CP(0, T) & CP(1, T) & \cdots & CP(c-1, T) \\ CP(0, T) & I + CP(1, T) & \cdots & CP(c-1, T) \\ \vdots & \vdots & \ddots & \vdots \\ CP(0, T) & CP(1, T) & \cdots & I + CP(c-1, T) \end{bmatrix} \quad (\text{H.21})$$

where

$$C = \left[I - \sum_{v=0}^{c-1} P(v, T) \right]^{-1}. \quad (\text{H.22})$$

14. Find the column vector :

$$\begin{aligned} \underline{k}^* &= \underline{e} + A_0(I - B_0)^{-1}\underline{e} \\ &+ \left\{ C_0(I - B_0)^{-1}(\underline{B}_3 - \underline{B}_1 + \underline{B}_2) + (\underline{A}_3 - \underline{A}_1 + \underline{A}_2) \right\} \cdot \\ &\left(I - \underline{G} + \underline{\tilde{G}} \right)^{-1} \underline{\phi}. \end{aligned} \quad (\text{H.23})$$

15. Find the boundary probability vectors :

$$\begin{aligned} \underline{\hat{x}}_0 &= \left(\underline{\hat{g}} \underline{\phi}_0 \right)^{-1} \left[\underline{\hat{g}} G(0, 0), \underline{\hat{g}} G(0, 1), \dots, \underline{\hat{g}} G(0, c-1) \right], \\ \underline{\hat{x}}_1 &= (\underline{k} \underline{k}^*)^{-1} \underline{k}. \end{aligned} \quad (\text{H.24})$$

16. Check $\underline{\hat{x}}_0 = \underline{\hat{x}}_1 A_0 + \underline{\hat{x}}_0 B_0$.

The above procedures are based on the reference [2, 25]. In order to find the probability vectors, $\underline{\hat{x}}_i$ ($i > 1$), the method by Ramaswami [10], which reduces the problem of round-off error propagation and also applicable when A_0 is singular, is employed.

17. Find the matrices S_i and $\tilde{S}_i$:

$$\begin{aligned} S_i &= \sum_{v=i+1}^{\infty} A_v G^{v-i-1} (I - A_1)^{-1}, \quad i \geq 1, \\ \tilde{S}_i &= \sum_{v=i}^{\infty} B_v G^{v-i} (I - A_1)^{-1}, \quad i \geq 2, \end{aligned} \quad (\text{H.25})$$

where the matrix S_i is the matrix whose entries are the expected number of visits to the state in level $j+i$ starting from a state in level j before hitting states in level k , where $k < j+i$ ($j > 0$), and $\tilde{S}_i$ is similarly defined except that the starting state is in level 0.

18. Find $\underline{x}_i$'s recursively by

$$\underline{\hat{x}}_{i+1} = \sum_{j=1}^i \underline{\hat{x}}_j S_{i+1-j} + \underline{\hat{x}}_0 \tilde{S}_{i+1}, \quad i \geq 1. \quad (\text{H.26})$$

Note : $\underline{\hat{x}}_i$'s can be truncated at any level. In the analysis, the truncation level is determined when $\underline{\hat{x}}_i$ is less than 10^{-8} .

Appendix I

Analysis of Queueing Performance for Low Priority Packets — Approximation 2 (ref. Chap. 5 — Sect. 5.3.2)

Following is the approximate analysis for the low priority packet queueing performance. This approximation is a further simplification of *approximation 1* (found in Sect. 5.3.2 Chapter 5) by aggregating the arrival state information for the high priority packet and only retaining the partial information of the high priority queue length.

Denote $\bar{q}$, $\bar{q} \geq c$ as the upper limit on the information of the high priority queue length that is to be available for the low priority packets; the information for high priority queue length that is beyond $\bar{q}$ is aggregated at the upper-bound $\bar{q}$. The low priority packet queueing system observed just prior to the transmission of the n^{th} slot is described *approximately* by the state vector

$$\underline{S}_n = (\underline{L}Q_n, {}_H\bar{Q}_n), \quad (\text{I.1})$$

where

$\underline{L}Q_n$ = the number of low priority packets waiting in the low priority queue

$${}_H\bar{Q}_n = \begin{cases} i, & 0 \leq {}_HQ_n < \bar{q} \\ \bar{q}, & {}_HQ_n \geq \bar{q} \end{cases} \quad (\text{I.2})$$

${}_HQ_n$ = the number of high priority packets waiting in the high priority queue

The sequence $\{\underline{L}S_n, n \geq 1\}$ is a 2-dimensional, first-order Markov chain and the transition probability matrix of the servers' availability is again obtained from $\mathbf{P}$ (Eqn.(5.16)) with the transition probabilities at $\bar{q}$ and beyond is aggregated at the boundary, i.e. at $\bar{q}$. Denote $\mathcal{T}$ as the duration of one slot (or packet) transmission and define

$$\hat{p}_{i,j} = \Pr \left[{}_H\tilde{Q}_n = j \mid {}_H\tilde{Q}_{n-1} = i \right], \quad n \geq 1, 0 \leq i, j \leq \bar{q}. \quad (1.3)$$

Then the availability matrix is given as

$$\mathbf{V} = \begin{bmatrix} \hat{p}_{0,0} & \hat{p}_{0,1} & \hat{p}_{0,2} & \cdots & \hat{p}_{0,\bar{q}-c} & \hat{p}_{0,\bar{q}-c+1} & \cdots & \hat{p}_{0,\bar{q}-1} & \hat{p}_{0,\bar{q}} \\ \hat{p}_{1,0} & \hat{p}_{1,1} & \hat{p}_{1,2} & \cdots & \hat{p}_{1,\bar{q}-c} & \hat{p}_{1,\bar{q}-c+1} & \cdots & \hat{p}_{1,\bar{q}-1} & \hat{p}_{1,\bar{q}} \\ \vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots \\ \hat{p}_{c-1,0} & \hat{p}_{c-1,1} & \hat{p}_{c-1,2} & \cdots & \hat{p}_{c-1,\bar{q}-c} & \hat{p}_{c-1,\bar{q}-c+1} & \cdots & \hat{p}_{c-1,\bar{q}-1} & \hat{p}_{c-1,\bar{q}} \\ \hat{p}_{c,0} & \hat{p}_{c,1} & \hat{p}_{c,2} & \cdots & \hat{p}_{c,\bar{q}-c} & \hat{p}_{c,\bar{q}-c+1} & \cdots & \hat{p}_{c,\bar{q}-1} & \hat{p}_{c,\bar{q}} \\ 0 & \hat{p}_{c+1,0} & \hat{p}_{c+1,1} & \cdots & \hat{p}_{c+1,\bar{q}-c-1} & \hat{p}_{c+1,\bar{q}-c} & \cdots & \hat{p}_{c+1,\bar{q}-2} & \hat{p}_{c+1,\bar{q}-1} \\ \vdots & \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \hat{p}_{\bar{q}-1,1} & \hat{p}_{\bar{q}-1,2} & \cdots & \hat{p}_{\bar{q}-1,c} & \hat{p}_{\bar{q}-1,c+1} \\ 0 & 0 & 0 & \cdots & \hat{p}_{\bar{q},0} & \hat{p}_{\bar{q},1} & \cdots & \hat{p}_{\bar{q},c-1} & \hat{p}_{\bar{q},c} \end{bmatrix} \quad (1.4)$$

where

$$\hat{p}_{i,j} = \begin{cases} (\underline{x}_i \underline{e})^{-1} \underline{x}_i \sum_{l=\bar{q}}^{\infty} P(l, \mathcal{T}) \underline{e} & \text{if } 0 \leq i \leq c, j = \bar{q} \\ (\underline{x}_i \underline{e})^{-1} \underline{x}_i \sum_{l=j}^{\infty} P(l, \mathcal{T}) \underline{e} & \text{if } c < i < \bar{q}, j = c + \bar{q} - i \\ \left(\sum_{l=\bar{q}}^{\infty} \underline{x}_l \underline{e} \right)^{-1} \sum_{l=\bar{q}}^{\bar{q}+j-1} \underline{x}_l P(\bar{q} + j - l, \mathcal{T}) \underline{e} & \text{if } i = \bar{q}, 0 \leq j < c \\ \left(\sum_{l=\bar{q}}^{\infty} \underline{x}_l \underline{e} \right)^{-1} \sum_{r=\bar{q}}^{\infty} \underline{x}_r \sum_{s=\max(\bar{q}+c-r, 0)}^{\infty} P(s, \mathcal{T}) \underline{e} & \text{if } i = \bar{q}, j = c \\ (\underline{x}_i \underline{e})^{-1} \underline{x}_i P(j, \mathcal{T}) \underline{e} & \text{if } (0 \leq i \leq c, j < \bar{q}) \text{ or} \\ & (c < i < \bar{q}, j < c + \bar{q} - i) \end{cases} \quad (1.5)$$

Define the joint steady state probability for low priority packet as

$$y_{i,j} = \Pr \left[{}_L Q_n = i, {}_H \tilde{Q}_n = j \right], \quad i \geq 0, \quad 0 \leq j \leq \tilde{q} \quad (1.6)$$

and the invariant probability vector for the queue length distribution of low priority packets as

$$\underline{y} = (\underline{\hat{y}}_0, \underline{\hat{y}}_1, \underline{\hat{y}}_2, \dots), \quad (1.7)$$

where the partitions are given as

$$\begin{aligned} \underline{\hat{y}}_k &= (\hat{y}_{kc}, \hat{y}_{kc+1}, \dots, \hat{y}_{(k+1)c-1}), \quad k \geq 0, \quad \text{and} \\ \underline{\hat{y}}_{kc+i} &= (y_{kc+i,0}, y_{kc+i,1}, \dots, y_{kc+i,\tilde{q}}), \quad 0 \leq i \leq c-1. \end{aligned} \quad (1.8)$$

Then the stochastic transition probability matrix for the low priority packet queueing system is given by

$$\tilde{\mathbf{P}} = \begin{bmatrix} \tilde{B}_0 & \tilde{B}_1 & \tilde{B}_2 & \tilde{B}_3 & \dots \\ \tilde{A}_0 & \tilde{A}_1 & \tilde{A}_2 & \tilde{A}_3 & \dots \\ 0 & \tilde{A}_0 & \tilde{A}_1 & \tilde{A}_2 & \dots \\ 0 & 0 & \tilde{A}_0 & \tilde{A}_1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \quad (1.9)$$

which satisfies

$$\underline{y} \tilde{\mathbf{P}} = \underline{y}, \quad \underline{y} \underline{e} = 1 \quad (1.10)$$

with the submatrices $\tilde{B}_i$'s, $\tilde{A}_i$'s, $i \geq 0$ are of order $c \times (\tilde{q} + 1)$ and they have the forms

given by

$$\begin{aligned} \hat{B}_i &= \begin{bmatrix} C_{0,ic} & C_{0,ic+1} & \cdots & C_{0,(i+1)c-1} \\ C_{1,ic} & C_{1,ic+1} & \cdots & C_{1,(i+1)c-1} \\ \vdots & \vdots & \ddots & \vdots \\ C_{c-1,ic} & C_{c-1,ic+1} & \cdots & C_{c-1,(i+1)c-1} \end{bmatrix}, \quad i \geq 0 \\ \tilde{A}_i &= \begin{bmatrix} C_{c,ic} & C_{c,ic+1} & \cdots & C_{c,(i+1)c-1} \\ C_{c,ic-1} & C_{c,ic} & \cdots & C_{c,(i+1)c-2} \\ \vdots & \vdots & \ddots & \vdots \\ C_{c,(i-1)c+1} & C_{c,(i-1)c+2} & \cdots & C_{c,ic} \end{bmatrix}, \quad i \geq 0 \end{aligned} \quad (1.11)$$

and

$$C_{r,s} = \Delta(\underline{k}) \cdot \mathbf{V}, \quad 0 \leq r \leq c, \quad s \geq 0 \quad (1.12)$$

where $\Delta(\underline{k})$ is a diagonal matrix of order $(\bar{q} + 1)$ with the diagonal elements form a vector $\underline{k}$ given by

$$\underline{k} = [k_0, k_1, \dots, k_{\bar{q}}] \quad (1.13)$$

where

$$k_{v,(0 \leq v \leq \bar{q})} = \begin{cases} \alpha_w & \text{if } w \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad (1.14)$$

with

$$w = r - \min(v, c) + \min(c - r, \min(v, c)), \quad \text{and} \quad (1.15)$$

$$\alpha_w = \text{Prob. of } w \text{ low priority packets arrive in time } T.$$

Since $\tilde{\mathbf{P}}$ indicates that the Markov chain is again of M/G/1 Type, the steady state probability distribution can be readily obtained following the procedure outlined in Appendix A.

3

of/de

3

MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS
STANDARD REFERENCE MATERIAL 1010a
(ANSI and ISO TEST CHART No. 2)

References

- [1] M. F. Neuts, *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. Baltimore, Maryland: The Johns Hopkins University Press, 1981.
- [2] D. Lucantoni and M. Neuts, "Numerical Methods for a Class of Markov Chains Arising in Queueing Theory." Tech. rept. no. 78/10. Applied Mathematics Institute, University of Delaware, Newark, 1978.
- [3] A. Konheim, "A Queueing Analysis of Two ARQ Protocols," *IEEE Trans. Commun.*, vol. COM-28, no. 7, pp. 1004-1014, July 1980.
- [4] M. Anagnostou and E. Protonotarios, "Performance Analysis of the Selective Repeat ARQ Protocol," *IEEE Trans. Commun.*, vol. COM-34, no. 2, pp. 127-135, Feb. 1986.
- [5] L. Golding, "Small Aperture Earth Station Networks and their relationship to ISDN," *Conf. Rec. IEEE Globecom 87*, pp. 54.4.1-54.4.4, 1987.
- [6] W. Wu and D. Tudge, "Values of ISDSN Attributes," *Conf. Rec. IEEE Globecom 87*, pp. 54.1.1-54.1.4, 1987.
- [7] T. Suda and N. Watanabe, "Evaluation of error recovery schemes for a high-speed packet switched network: link-by-link versus edge-to-edge schemes," *Proc. IEEE INFOCOM 1988*, vol. 1, pp. 722-731, 1988.
- [8] A. Bhargava, J.F. Kurose, D. Towsiey and G. Leemput, "Performance comparison of error control schemes in high speed computer communication networks," *Proc. IEEE INFOCOM 1988*, vol. 1, pp. 694-703, 1988.
- [9] H. Burton and D. Sullivan, "Errors and Error Control," *Proc. of IEEE*, vol. 60, no. 11, pp. 1293-1301, Nov. 1972.

- [10] V. Ramaswami, "A Stable Recursion for the Steady State Vector in Markov Chains of M/G/1 Type," *Commun. Statist. - Stochastic Models*, vol. 4(1), pp. 183-188, 1988.
- [11] J. Metzner and D. Chang, "Efficient Selective Repeat ARQ Strategies for Very Noisy and Fluctuating Channels," *IEEE Trans. Commun.*, vol. COM-33, no. 5, pp. 409-415, May 1985.
- [12] L.R. Lugand, D.J. Costello, Jr. and R. Deng, "Parity Retransmission Hybrid ARQ Using Rate 1/2 Convolutional Codes on a Nonstationary Channel," *IEEE Trans. Commun.*, vol. 37, no. 7, pp. 755-765, July 1989.
- [13] X.Wang and S.Chang, "On the Performance Study of Several Access Control Strategies in ISDN," in *Conf. Rec. IEEE ICC 88*, vol. 1, pp. 934-938, 1988.
- [14] Y. DeSerres and L. Mason, "A Multiserver Queue with Narrow- and Wide-Band Customers and Wide-Band Restricted Access," *IEEE Trans. Commun.*, vol. 36, no. 6, pp. 675-684, Jun. 1988.
- [15] K. Liao and L. Mason, "An Approximate Performance Model for a Multislot Integrated Services System," *IEEE Trans. Commun.*, vol. 37, no. 3, pp. 211-221, Mar. 1989.
- [16] B. Kraimeche and M. Schwartz, "Analysis of Traffic Access Control Strategies in Integrated Service Networks," *IEEE Trans. Commun.*, vol. COM-33, no. 10, pp. 1085-1093, Oct. 1985.
- [17] B. Kraimeche and M. Schwartz, "A Channel Access Structure for Wideband ISDN," *IEEE J. Select. Areas Commun.*, vol. SAC-5, no. 8, pp. 1327-1335, Oct. 1987.
- [18] B. Kraimeche and M. Schwartz, "Bandwidth allocation strategies in Wide-Band Integrated Networks," *IEEE J. Select. Areas Commun.*, vol. SAC-4, no. 6, pp. 869-878, Sept. 1986.

- [19] Y. Kim and C. Un, "Bandwidth allocation strategy with access restriction and preemptive priority," *Electronics Letters*, vol. 25, no. 10, pp. 655-656, May 1989.
- [20] D. Miller, "Steady-State Algorithmic Analysis of M/M/c Two-Priority Queues with Heterogeneous Rates," in *Applied Probability - Computer Science : The Interface Vol. II* (R. Disney and T. Ott, eds.), pp. 207-225, Birkhauser Boston Inc., 1982.
- [21] D. Miller, "Computation of Steady-State Probabilities for M/M/1 Priority Queues," *Operations Research*, vol. 29, no. 5, pp. 945-957, Sept.-Oct. 1981.
- [22] B. Maglaris, D. Anastassiou, P. Sen, G. Karlsson, and J. Robbins, "Performance Models of Statistical Multiplexing in Packet Video Communications," *IEEE Trans. Commun.*, vol. 36, no. 7, pp. 834-843, July 1988.
- [23] H. Heffes and D. Lucantoni, "A Markov Modulated Characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Performance," *IEEE J. Select. Areas Commun.*, vol. SAC-4, no. 6, pp. 856-867, Sept. 1986.
- [24] M. Neuts, "A Versatile Markovian Point Process," *J. Appl. Prob.*, vol. 16, pp. 764-779, Dec. 1979.
- [25] M. Neuts, "The c-server Queue with Constant Service Times and a Versatile Markovian Arrival Process," in *Applied Probability - Computer Science : The Interface Vol. II* (R. Disney and T. Ott, eds.), pp. 31-70, Birkhauser Boston Inc., 1982.
- [26] P. Sen, B. Maglaris, N.E. Rikli and D. Anastassiou, "Models for Packet Switching of Variable-Bit-Rate Video Sources," *IEEE J. Select. Areas Commun.*, vol. 7, no. 5, pp. 865-869, June 1989.
- [27] J. Mathews, *Numerical Methods : for computer science, engineering, and mathematics*. London, England: Prentice-Hall International, Inc., 1987.

- [28] V. Ramaswami, "Nonlinear Matrix Equations in Applied Probability Solution Techniques and Open Problems," *SIAM Review*, vol. 30, no. 2, pp. 256–263, June 1988.
- [29] D. Jagerman, "An Inversion Technique for the Laplace Transform," *Bell Syst. Tech. J.*, vol. 61, no. 8, pp. 1995–2002, Oct. 1982.

END

1|2|0|4|9|1