

ASYMPTOTIC ANALYSIS OF HIGHLY RELIABLE  
RETRIAL QUEUEING SYSTEMS

A THESIS  
SUBMITTED TO THE DEPARTMENT OF INDUSTRIAL  
ENGINEERING  
AND THE INSTITUTE OF ENGINEERING AND SCIENCES  
OF BILKENT UNIVERSITY  
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR THE DEGREE OF  
MASTER OF SCIENCE

By  
Mümin Kurbanlı  
June, 2000

THESIS

T

57.9

K87

2000

ASYMPTOTIC ANALYSIS OF HIGHLY RELIABLE  
RETRIAL QUEUEING SYSTEMS

A THESIS  
SUBMITTED TO THE DEPARTMENT OF INDUSTRIAL  
ENGINEERING  
AND THE INSTITUTE OF ENGINEERING AND SCIENCES  
OF BILKENT UNIVERSITY  
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR THE DEGREE OF  
MASTER OF SCIENCE

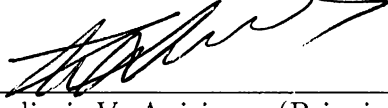
By  
Mümin Kurtuluş  
June, 2000



B052747

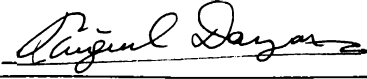
T  
57.9  
.K87  
2000

I certify that I have read this thesis and that in my opinion it is fully adequate,  
in scope and in quality, as a thesis for the degree of Master of Science.



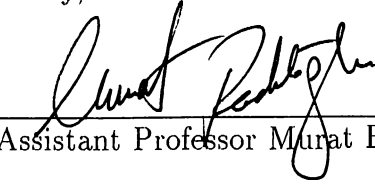
\_\_\_\_\_  
Professor Vladimir V. Anisimov (Principal Advisor)

I certify that I have read this thesis and that in my opinion it is fully adequate,  
in scope and in quality, as a thesis for the degree of Master of Science.



\_\_\_\_\_  
Assistant Professor Tuğrul Dayar

I certify that I have read this thesis and that in my opinion it is fully adequate,  
in scope and in quality, as a thesis for the degree of Master of Science.



\_\_\_\_\_  
Assistant Professor Murat Fadiloğlu

Approved for the Institute of Engineering and Sciences:



\_\_\_\_\_  
Prof. Mehmet Baray  
Director of Institute of Engineering and Sciences

# ABSTRACT

## ASYMPTOTIC ANALYSIS OF HIGHLY RELIABLE RETRIAL QUEUEING SYSTEMS

Mümin Kurtuluş

M.S. in Industrial Engineering

Supervisor: Professor Vladimir V. Anisimov

June, 2000

The thesis is concerned with the asymptotic analysis of the time of first loss of a customer and the flow of lost customers in some types of Markov retrial queueing systems with finite buffer. A retrial queueing system is characterized by the following feature: an arriving customer finding all of the servers busy must leave the service area and join a special buffer. After this it may re-apply for service after some random time. If the buffer is full the customer is lost. The analysis of the time of first loss of a customer is based on the method of so-called  $S$  – sets and the results about the asymptotic behavior of the first exit time from the fixed subset of states of semi-Markov process of a special structure (so-called *monotone structure*). Single server retrial queueing systems ( $M/M/1/m$  with retrials) as well as multiple server retrial queueing systems ( $M/M/s/m$  with retrials) are analyzed in cases of fast service and both fast service and fast retrials. Exponential approximation for the time of first loss and Poisson approximation for the flow of lost customers are proved for all of the considered cases.

*Keywords:* Retrial queueing systems, rare events, s-set, asymptotic analysis.

## ÖZET

### ÇOK GÜVENİLİR TEKRAR DENEMELİ SIRA SİSTEMLERİNİN ASİMTOTİK ANALİZİ

Mümin Kurtuluş

Endüstri Mühendisliği Yüksek Lisans

Tez Yöneticisi: Profesör Vladimir V. Anisimov

Haziran, 2000

Bu tez çalışması, bazı Markov tekrar denemeli sıra sistemlerindeki ilk müşteriyi kaybetme zamanının asimtotik analizi ile ilgilidir. Tekrar denemeli sıra sistemleri şu özellikleri ile tanımlanabilir: sisteme girdiği anda, bütün makinaları meşgul bulan müşteri, makinaların bulunduğu alanı terkeder ve özel bir sıraya dahil olur. Müşteri rassal bir süreç sonunda tekrar makinalara servis için başvurabilir. Geldiğinde, bütün makinaları ve bekleme yerlerini dolu bulan müşteriler, sistemden uzaklaşır (kaybedilir). Birinci müşteriyi kaybetme zamanı ile ilgili analiz, S-kümeleri diye bilinen ve monoton (tekdüze) yapılar diye bilinen kavramlar yardımı sayesinde yapılmıştır. Bir-makineli ve birden fazla makineye sahip olan sistemler için analiz yapılmıştır. Makinaların süratli çalıştığı ve tekrar servis için deneyen müşterilerin bunu hızlı yaptığı farz edilerek, yukarıda sözü edilen sistemler incelenmiştir. Birinci kaybedilen müşterinin zaman dağılımının exponential olduğu kanıtlanmıştır.

*Anahtar sözcükler.* Tekrar denemeli sıra sistemleri, nadir olaylar, s-kümeleri, monoton yapılar, asimtotik analiz.

*To my family,*

## ACKNOWLEDGEMENT

I would like to express my sincere gratitude to Professor Vladimir V. Anisimov for his invaluable advice and supervision in this thesis work. I learned a lot from him. He has been guiding me with patience and everlasting interest not only for this research but also for my future career.

I am indebted to Assistant Profesor Tuğrul Dayar and Assistant Professor Murat Fadiloğlu for showing keen interest in the subject matter and accepting to read and review this thesis.

I am grateful to Professor Lee Schruben for showing keen interest to the problem and helping me in simulation of the problem.

I am also grateful to Ayten Türkcan for helping me in preparation of the  $\LaTeX$  draft of this thesis.

I would like to thank my officemates Bilal Ayduran, Ahmet Reha Botsalı, Tolga Genç and Batuhan Kızılişik who were always helpful and understanding. We shared a lot in the last two years.

I also wish to thank Huseyin Karakuş, Yasin Dağistan, Umut Çetin, Selçuk Onay, Ali Irek and Hürer Fethi Gündüz for being such a good friends during the last six years of my life. Without their friendship and support, I would not be able to bear with all this time.



# Contents

<b>1</b>	<b>INTRODUCTION AND THE GENERAL MODEL</b>	<b>1</b>
1.1	The General Retrial Queueing Model . . . . .	2
1.2	Examples of retrial queueing systems . . . . .	6
<b>2</b>	<b>LITERATURE REVIEW</b>	<b>8</b>
2.1	Retrial queues . . . . .	8
2.1.1	Single-server retrial queues . . . . .	9
2.1.2	Multiple-server retrial queues	10
2.1.3	Retrial queues with waiting positions . . . . .	11
2.1.4	Retrial queues with batch arrivals . . . . .	13
2.2	Asymptotic analysis of rare events in queueing models . . . . .	14
<b>3</b>	<b>PRELIMINARY RESULTS</b>	<b>18</b>
3.1	Results about asymptotic behavior of the first exit time from the fixed subset of states of SMP . . . . .	18

3.2	$M/M/s/m$ queueing system . . . . .	25
<b>4</b>	<b>SINGLE-SERVER RETRIAL QUEUEING MODELS</b>	<b>27</b>
4.1	$M/M/1/m$ system with retrials . . . . .	27
4.2	$M_U/M_U/1/m$ retrial queueing system operating in Markov environment . . . . .	36
<b>5</b>	<b>MULTIPLE-SERVER RETRIAL QUEUEING MODELS</b>	<b>41</b>
5.1	$M/M/2/m$ system with retrials . . . . .	42
5.2	$M/M/s/m$ system with retrials . . . . .	48
<b>6</b>	<b>SIMULATION RESULTS</b>	<b>53</b>
6.1	Simulation of $M/M/1/m$ system with retrials . . . . .	54
6.2	Simulation of $M/G/1/m$ system with retrials . . . . .	57
<b>7</b>	<b>CONCLUSION</b>	<b>61</b>
	Vita	72

# List of Figures

1.1	General Retrial Queueing Model . . . . .	3
1.2	Retrial Queueing Model where repeated calls originate from the waiting positions . . . . .	4
3.1	S-set . . . . .	19
3.2	Monotone Structure . . . . .	22
4.1	Monotone structure for single server model with assumption of fast service . . . . .	31
4.2	Monotone structure for the model with single server and assumptions of fast service and fast retrials . . . . .	34
4.3	Monotone structure for single server system which operates in additional Markov environment and assumption of fast service . . . . .	38
5.1	Monotone structure for the model with two servers and assumption of fast service . . . . .	44
5.2	Monotone structure for the model with two servers and assumptions of fast service and fast retrials . . . . .	47

6.1	Approximated and simulated densities for the time of loss of first customer in a $M/M/1/m$ system with retrials where $m = 2$ , $\lambda = 1$ , $\mu = 10$ $\nu = 2$ .	56
6.2	Approximated and simulated densities for the time of loss of first customer in a $M/G/1/m$ with retrials where $m = 2$ , $\lambda = 1$ , $\nu = 2$ and service times are uniformly distributed on the interval $[0, 0.5]$ .	58
6.3	Approximated and simulated densities for the time of loss of first customer in a $M/G/1/m$ with retrials where $m = 2$ , $\lambda = 1$ , $\nu = 2$ and service times are uniformly distributed on the interval $[0, 0.2]$ .	59
7.1	Summary of the results for single server retrial queueing models	64
7.2	Summary of the results for multiple server retrial queueing models	65

# List of Tables

- 6.1 Results of simulation of the time of first customer loss in the system of the type  $M/M/1/m$  with retrials where  $m = 2$ ,  $\lambda = 1$ ,  $\mu = 10$ , and  $\nu = 2$  . . . . . 55
  
- 6.2 Results of simulation of the time of first customer loss in the system of the type  $M/G/1/m$  with retrials where  $m = 2$ ,  $\lambda = 1$ ,  $\nu = 2$  and service times are uniformly distributed on the interval  $[0, 0.5]$ . . . . . 60
  
- 6.3 Results of simulation of the time of first customer loss in the system of the type  $M/G/1/m$  with retrials where  $m = 2$ ,  $\lambda = 1$ ,  $\nu = 2$  and service times are uniformly distributed on the interval  $[0, 0.2]$ . . . . . 60

# Chapter 1

## INTRODUCTION AND THE GENERAL MODEL

Real life mathematical models of computing systems, telephone switching systems and communication networks usually have complex hierarchical structure and operate in different scales of time. For example real time and computer time are in different scales. Even for Markov models, exact analytic solutions can be obtained only for special rare cases. Therefore asymptotic methods and approximation techniques play a very important role in investigation and modeling of such systems.

In many models of practical interest, usually “small parameters” are present, e.g., the rate of incoming customers in a system is much smaller than the rate of service (in queueing theory this is termed as “fast service”). These small parameters give rise to the so called flows of rare events in reliability and queueing theory. In applications a rare event usually means different types of failures, an exit from some region, a loss of call, exceeding some level, etc.

The thesis is devoted to the asymptotic analysis of stochastic systems with finite number of states and different orders of transition probabilities. Analysis is oriented toward Markov retrial queueing system with finite buffer and which operates in different scales of time.



Queueing systems in which arriving customers who find all servers and waiting positions (if any) occupied may retry for service after a period of time are called retrial queues or queues with repeated customers.

In the simplest and best known queueing system models, an arriving customer may receive immediate service, may wait in line to receive immediate service after some future departure, or may leave the system without receiving service. Retrial queueing models attempt to capture a property of many real queueing systems not present in the simple models - that a customer not receiving immediate service on arrival may return at a later time to try again. A customer waiting to return is said to be in orbit. In a system of this type a server may be idle while unserved customers remain in the system.

The theory of retrial queues, like queueing theory itself, had its origin in problems of communication. Retrial queues have been widely used to model many problems in telephone switching systems, telecommunication networks, computer networks and computer systems.

## 1.1 The General Retrial Queueing Model

The general retrial queueing model consists of  $s$  identical independent servers and  $m$  waiting positions. Customers arrive at the system according to a Poisson process with parameter  $\lambda$ . The service time for each customer served is an independent identically distributed random variable. On arrival, if one or more of the servers are free, the customer will receive service immediately; otherwise, if none of the servers is free and there are free waiting positions, the customer will join the queue waiting for service. On the other hand, if an incoming customer finds all servers and waiting positions full, the customer will leave the system forever with probability  $1 - H_0$  or leave the service area temporarily with probability  $H_0$  and will retry for service after a random period of time. Those customers who will retry for service are said to be "in orbit". The capacity of orbit is denoted as  $O$  and can be either finite or infinite. If the

orbit is full in the case of finite  $O$ , any customer coming to the orbit will be forced to leave the system forever. Each orbiting customer will retry for service with independent input rate of  $\nu$ . Customers retrying for service are treated as primary customers. Again if customer finds free server he will start service immediately or if the servers are full and there is free waiting position, he will join the queue waiting for service. On the other hand, if an incoming customer finds all servers and waiting positions full, customer will leave the system forever with probability  $1 - H_k$  (if it is the  $k^{\text{th}}$  unsuccessful retrial) or join the orbit, if the orbit is not full, with probability  $H_k$ .

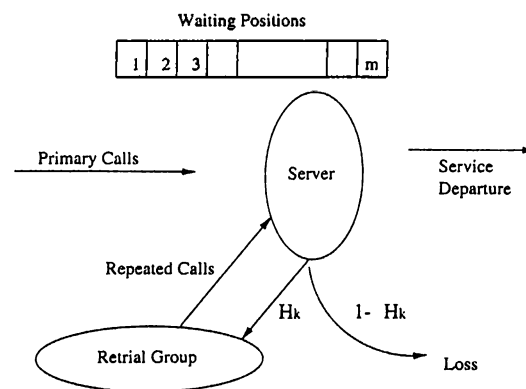


Figure 1.1: General Retrial Queueing Model

The model described above is quite general one and many of the retrial queueing systems can be considered as its special case. An extended Kendall notation of the form  $A/B/s/m/O/H$  can be used to represent a general retrial queueing system. Retrial time is not described in the notation above. Usually the retrial time is exponential with parameter  $\nu$ .

- $A$  Describes interarrival time distribution.
- $B$  Describes the service time distribution.
- $s$  The number of servers in the system.
- $m$  The number of waiting positions in the system.
- $O$  Capacity of the orbit.
- $H$  Stands for the loss model and can be described as a series  $H_0, H_1, H_2, \dots$

When  $H_k = 1$  for  $k \geq 0$ , every customer receives service if  $O$  is infinite and such systems are called no-loss systems ( $H$  is NL). On the other hand, when  $H_k = \alpha < 1$  for  $k \geq 0$  the system is called a geometric loss system ( $H$  is GL).

As stated previously, the model described above is quite general and many of the models considered previously are special cases of this model. The model that we considered through the thesis can be described as follows.

Consider the system with single server and  $m$  waiting places. The customers arrive to the system one at a time. Upon arrival, if the server is free, the customers receive service immediately, otherwise the customer joins the special queue which we call retrial queue. The customers waiting in the retrial queue will attempt for service after some random period of time. If a retrying customer finds the server empty, he will receive service immediately, otherwise (if the server is busy) he will return to the retrial queue and retry for service later. If an incoming customer finds the server and all of the waiting positions occupied, the customer will leave the system forever.

Although the model considered in this thesis is a retrial queueing system with waiting positions, it is a special form of the model considered in the literature. Unlike the models considered in the literature, we assume that repeated calls originate from waiting positions (i.e, customers rejected from service form a special queue from which they repeat their attempt for service).

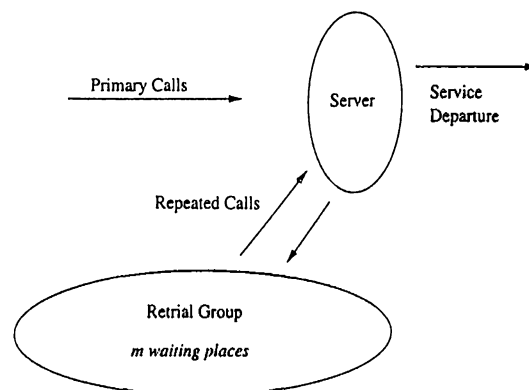


Figure 1.2: Retrial Queueing Model where repeated calls originate from the waiting positions

We will consider the time of loss of first customer for single server and multiple server retrial queues under some different assumptions.

Chapter 2 is divided into two parts. The first part is devoted to the literature review on retrial queues. We consider retrial queues of single server and multiple server types as well as retrial queues with waiting positions and retrial queues with batch arrivals which are the models of interest in the area. The second part is devoted to literature review on asymptotic analysis of rare events in queueing systems which essentially forms a basis for this thesis.

In Chapter 3, the results about asymptotic behavior of first exit time from a fixed subset of states of a SMP are reviewed. The exponential approximation for the time of exit is proved. Notions of  $S$  – set and monotone structure are introduced.

Chapter 4 deals with the time of loss of first customer in an  $M/M/1/m$  queueing system with retrials under some different assumptions. The method of analysis is based on the results about asymptotic behavior of the first exit time from the fixed subset of states of a SMP which forms an  $S$  – set, which we give in chapter 3. First we consider asymptotic behavior of the system under the assumption of fast service and then we consider the system under the assumptions of both fast service and fast retrials. Finally, we consider the system operating in Markov environment and under the assumption of fast service. We derive the expression for the parameter of exponential distribution for the time of loss of first customer for these models.

In Chapter 5, we study the multiple-server retrial queueing system of type  $M/M/s/m$  and derive the asymptotic expression for the parameter of exponential distribution for the time of loss of first customer under the assumptions of fast service and fast retrials. The method of derivation is as follows: first we study the system with only 2 servers ( $s = 2$ ) and  $m$  waiting places, then the general result for  $s$  server case is obtained by generalizing the previous results.

Chapter 6 is devoted to the simulation of retrial queueing systems. Time of

first loss of customer in Markov retrial queueing system is simulated and results of simulation and approximate results are compared. Also time of first loss of customer in retrial queueing system where service is assumed to be uniformly distributed is simulated and results are compared. The service is assumed to be fast for both cases.

## 1.2 Examples of retrial queueing systems

Retrial queues arise naturally as models of many problems in telecommunication, computer networks and computer systems, and in daily life. In this subsection, we give some examples of problems which can be modeled as retrial queues.

1. *Making reservations.* Consider a service shop in which most of the reservations are made through telephone calls. There is only one line which is dedicated to answering requests for reservations. Normally, if a customer calls the service shop and finds the line busy, the customer tries the number again after some random period of time with probability  $H_k$  ( $H_k < 1$ ) if it is the  $k^{\text{th}}$  unsuccessful retrial. This example can be modeled as an  $M/G/1$  retrial queue with loss if the arrival process is approximated as Poisson. However, when  $H_k \rightarrow 1$  it can be approximated as an  $M/G/1$  retrial queue without loss. The interesting questions about the model described above can be: How long will the busy period last? What is the average waiting time of a customer? How many customers will the service shop lose due to blocking?
2. *Real time computer system.* Consider a real time computer system in which there are  $s$  ports and  $m$  ( $m \geq s$ ) terminals. For a terminal to be connected to the computer, exactly one port must be used. Students arrive at the computing center to use the computer for a random period of time. An arriving student must first find a free terminal to log on. If there is no free terminal, the student will normally try his luck after

some random time. If, on arrival, the student finds a free terminal, he will send commands to a central switch to request connection to the computer; otherwise the request will be queued by the switch and the student has to wait until there is a free port for him. This example can be modeled as multiple-server retrial queue with waiting positions (if  $m > s$ ) and infinite orbit capacity.

3. *Cellular phone subscriber.* Consider a cellular phone system. It is well known that a telephone subscriber who obtains a busy signal usually repeats the call until the required connection is made. As a result, the flow of calls circulating in a telephone network consists of two parts: the flow of primary calls, which reflects the real wishes of the telephone subscribers, and the flow of repeated calls, which is the consequence of the lack of success of previous attempts. If the subscriber finds telephone system available, he will be served, on the other hand cellular phone systems have an option that allows a second call to wait until the primary call is served. The system can be modeled as a single server retrial queueing system with single waiting place and system with losses. The interesting question to answer for the above system can be: What is the time of loss of first customer due to blocking?



# Chapter 2

## LITERATURE REVIEW

The pioneering work on the theory of queues was done by A. K. Erlang of the Copenhagen Telephone Company during 1909 to 1920. A systematic treatment of the theory from the point of view of stochastic processes is due to D. G. Kendall and this has greatly influenced subsequent works in this field. Many books and research papers were devoted to the theory of queues since that time [Prabhu [46], Neuts [45], Lipsky [44], etc.].

Queueing theory arise with the problems in telecommunication, therefore, the need for more realistic models give rise to the retrial queueing models. At present, the theory of retrial queues is recognized as an important part of queueing theory and teletraffic theory.

This chapter is devoted to the literature review on retrial queueing systems and includes a part which is devoted to the review of some asymptotic techniques used in analysis and investigation of queueing models.

### 2.1 Retrial queues

Retrial queueing models arise since the early works of Kosten[39], Cohen[24], Wilkinson[57], and Riordan[48]. Various techniques and results have been

developed to solve particular problems and to understand the nature of retrial queueing models. Some textbooks and monographs on queueing theory and teletraffic theory include sections devoted to retrial queues where only simple results concerning this type of systems are stated. The book by Falin and Templeton [33] is the first to consider retrial queues in full detail. Also, a detailed discussion of results obtained can be found in reviews by Yang and Templeton[58] and Falin[34].

In particular, the nature of results obtained, methods of analysis and areas of application allow us to divide retrial queues into two large groups: single-server systems and multiple-server systems.

### 2.1.1 Single-server retrial queues

The first result on  $M/G/1$  retrial queues is due to Keilson, Cozzolino and Young in [37] who used the method of supplementary variables. Most of the previous papers considered  $M/M/s$  retrial queues, where the  $M/M/1$  retrial queues are treated as a special case and both analytic and numerical results were obtained.

The system will be said to be in state  $(m, E)$  if the server is idle and  $m$  customers are orbiting. The system will be said to be in state  $(m, x)$  if the customer in service has been in service for time  $x$  and  $m$  customers are orbiting. The states  $\{(m, E), 0 \leq m < \infty\}$  and  $\{(m, x), 0 \leq m < \infty, 0 \leq x < \infty\}$  form the set of states of a Markov process. Using this method, the ergodic solutions are obtained for the generating functions of number of customers in the queue, as well as mean number of customers waiting, mean waiting time of a customer, and mean number of retrials per customer.

Later in [1] Aleksandrov studied the same model and obtained similar results using residual service time as a supplementary variable rather than elapsed service time and briefly discussed the structure of the busy period. The busy period is defined as the period that starts at the epoch when a call

enters an empty system and ends at the departure epoch at which the system is empty. The busy period consists of alternating service periods, and periods during which the server is free and there are sources in the system.

Choo and Conolly in [23] and Falin in [26] also examined the  $M/G/1$  retrial model and obtained some analytic results about the distribution of waiting time, system busy period, system idle time, system output flow, and the number of orbiting customers.

### 2.1.2 Multiple-server retrial queues

Multiple-server retrial queueing models have important applications in telephone switching systems. The earliest investigations in this area are by Kosten in [39], Wilkinson in [57], Cohen in [24] and Riordan in [48] where exponential interarrival time, exponential service time, and exponential retrial time ( $M/M/s$  retrial queues) were considered. The cases of finite and infinite orbit capacities and the possibility of lost customers were investigated. Steady state equations, major probabilistic characteristics of the system and analytic solutions for some special cases were obtained in these papers.

The later papers dealing with multiple server retrial queues can be viewed in two categories.

#### 2.1.2.1 Full-available systems

Multiple server retrial queueing system in which any idle server can be immediately seized by a primary or orbiting customer. Cohen in [24] studied  $M/M/s$  retrial queue of this type extensively.

Later Jonin and Sedol in [36] studied  $M/M/s$  retrial queue with  $H_k = \alpha \leq 1$  and derived explicit expressions for steady state probability that there are  $i$  busy servers and  $j$  orbiting customers in the system. Solving these equations is extremely difficult even for some special cases. Therefore, approximation

methods and asymptotic formulation are preferred in solving these problems.

Stepanov in [52] and Falin in [28] considered the method of asymptotic formulation. Stepanov in [52] considered  $M/M/s$  retrial queue with  $H_0 \leq 1$  and  $H_k = 1 (k \geq 1)$ . Asymptotic formulation of the model in the case of extreme load are presented for the system. Characteristics such as the blocking probability, the mean number of busy servers, and the mean number of orbiting customers are obtained.

Le Gall in [43] studied the  $M/G/s$  retrial queueing model with  $H_k = \alpha \leq 1$ , ( $k \geq 0$ ). Le Gall considered the blocking probability for primary customers, the mean blocking probability for the orbiting customers and the mean blocking probability for the orbiting customers.

### 2.1.2.2 Non-full-available systems

The system receives  $m$  independent Poisson flows of primary customers at rates  $\lambda_i$  ( $i = 1, 2, \dots, m$ ). Customers of the  $i^{th}$  flow can only choose one of  $V_i$  servers ( $V_1 + V_2 + \dots + V_m = s$ ) for servicing. If a customer from the  $i^{th}$  flow finds that all the servers that he can take are busy, then with probability  $H_{i,k}$  (if it is the  $k^{th}$  retrial) the customer will retry for service after an exponential amount of time.

Non-full-available systems are extensively studied by Falin in [27], and Stepanov in [50], [51], and [53]. Steady state equations are determined, and asymptotic formulations of the model are obtained for the major probabilistic characteristics of the system.

### 2.1.3 Retrial queues with waiting positions

Retrial queues with waiting positions occur in practical applications. Waiting positions, in many computer communication networks, telephone ordering systems and computer operating systems are frequently used to improve the

efficiency of the servers and to decrease the influence of the retrial customers. Research done on retrial queues with waiting positions can be classified in two groups as single server and multiple server retrial queues with waiting positions.

### 2.1.3.1 Single server retrial queue with waiting positions

Hashida and Kawashima in [35] considered the single server system with waiting positions and customer retrials where they assumed a geometric loss model with finite orbiting capacity. The model can be classified as  $M/M/1/m/O/GL$  with finite  $m$  and  $O$ . They also assume that  $H_k = \alpha$  for  $k \geq 0$ . The states are defined as  $(j, k)$  where  $j$  denotes the number of customers in the waiting room (including the one in service) and  $k$  denotes the number of customers orbiting. The authors derive the steady-state equations for the state probabilities and develop an efficient procedure to calculate the exact values of these state probabilities. Also, performance measures such as the mean queue length, mean number in the orbit, and mean waiting time are all expressed in terms of the state probabilities.

Later in [47] Ridout considered a different model where the orbit capacity is infinite (no-loss model) that is  $O = \infty$  and  $H_k = 1$  for all  $k \geq 0$ . The model analyzed by Ridout can be classified as  $M/M/1/m/O/NL$ . The states of the system are defined as  $(j, k)$ , as in the previous model. The steady-state equations for the state probabilities are derived but since both the number of unknowns and the number of equations are countable infinite, it is not easy to solve these equations for general  $m$ .

Ridout has developed analytical expressions for state probabilities in the case  $m = 1$ , and recursive procedures to calculate state probabilities for  $m = 2$  using generating functions. For  $m > 2$ , an iterative procedure was used to find approximate values of state probabilities.

### 2.1.3.2 Multiple server retrial queue with waiting positions

Stepanov and Tsitovich in [54] have studied a multiple server retrial queueing system with waiting positions which can be classified as  $M/M/s/m/O/NL$  with  $O = \infty$  and  $H_k = 1$  for  $k \geq 0$ . Although  $H_k = 1$ , their model is viewed as a loss system since they assumed that if a waiting customer does not succeed in receiving service, the customer leaves the waiting position after a random period which is exponentially distributed. The states of the system are defined as  $(j, i)$  where  $j$  is the number of orbiting customers and  $i$  is the number of customers in the waiting room or in service. The author consider the basic probabilistic characteristics such as the probability that all servers and waiting positions are busy and the distribution of mean number of orbiting customers. These quantities are considered in case of extreme load.

### 2.1.4 Retrial queues with batch arrivals

Retrial queues with batch arrivals are quite common in computer communication networks. In batch arrival retrial queues it is assumed that at every arrival epoch a batch of  $k$  primary calls arrive with probability  $c_k$ . If the channel is busy at the arrival epoch, then all these calls join the queue. On the other hand, if the channel is free, then one of the arriving customers begins service and the others form sources of repeated calls.

Falin in [25] considered the  $M/G/1$  retrial queue model with batch arrivals and no customer loss and obtained the probability generating function of the number of customers in the system. Falin also used the embedded Markov chain technique to derive the joint distribution of the channel state and the queue length.

Kulkarni in [42], has examined the same model but with two types of customers. Kulkarni obtained analytic expressions for the mean number of type  $i$  ( $i = 1, 2$ ) customers in the system, mean waiting time and mean number of retrials of a type  $i$  customer.



Models with multiclass (there are  $n$  types of customers and type  $i$  primary customers arrive in a Poisson process with rate  $\lambda_i$  and the retrial intensity is  $\nu_i$ ) batch arrivals were considered by Falin in [32] and Kulkarni in [42].

## 2.2 Asymptotic analysis of rare events in queueing models

Different asymptotic approaches for reliability analysis of various classes of stochastic systems are studied in the books [Borovkov [22], Korolyuk and Turbin [38], Kovalenko [40], Anisimov et al. [9], Anisimov [11]]. A survey of results devoted to the analysis of rare events in queueing systems is given by Kovalenko [41].

Anisimov in [2] introduced the concept of the so-called  $S$ -sets (asymptotically connected set). Several results devoted to the asymptotic analysis of integral functionals and flows of rare events on trajectories of the process with discrete component are obtained by Anisimov in [2], [4], [11]. The method of  $S$ -sets allows us to study the asymptotic behavior of the time of the first loss of a call for wide classes of Markov and semi-Markov processes with finite number of states and in case of fast service or light loading. Various applications of method of  $S$ -sets can be found in Anisimov et al. [9], Anisimov and Sztrik [12], [13], [14], Sztrik and Kouvatso [55], Sztrik [56] and Anisimov [16].

Anisimov and Sztrik [13] considered asymptotic analysis of a complex renewable system operating in random environment. Supposing "fast repair" it is shown that the time up to the first system failure converges in distribution, under appropriate normalization, to an exponentially distributed random variable. The failure and repair intensities of the elements depend on the indices of the failed elements and the state of the given random environment. This assumptions make the problem difficult. Using the results about method of  $S$  – sets and monotone structure the asymptotic exponentiality is proved.

Sztrik and Kouvatso [55] proposed an asymptotic queueing theoretic approach to analyze the performance of a FCFS (first come, first served) heterogeneous multiprocessor computer system with a single bus operating in a randomly changing environment. All stochastic times in the system are considered to be exponentially distributed and independent of the random environment, while the access and service rates of the processors are subject to random fluctuations. It is shown under the assumption of “fast” arrivals that the busy period length of the bus converges weakly, under appropriate normalization, to an exponentially distributed random variable. The results about  $S$  – sets were used in the proof.

Anisimov in [16], studied a Markov queueing system of the type  $M_M/M/\bar{l}/m$  where customers arrive according to a Poisson process where the local intensity of entry at time  $t$  is  $\lambda_i$  if  $x(t) = i$ . Here,  $x(t)$ ,  $t > 0$  is a continuous time Markov Process with finite state space  $\{1, 2, \dots, r\}$  given by the intensities of transition  $a_{ij}$ ,  $i = \overline{1, r}$ ,  $j = \overline{1, r}$ ,  $i \neq j$ . The system has  $l$  labeled servers and server  $i$  has intensity of service  $\mu_n(i)$ ,  $i = \overline{1, l}$ . The system has also  $m$  waiting places. The call entering the system occupies the free server with minimal label or joins the queue, if all of the servers are busy. If all servers and all waiting places are busy the call is lost. Supposing that the service is asymptotically fast, the author proved that time of first customer loss is exponentially distributed.

The operation of a wide class of stochastic systems can be described in terms of random processes such that the character of their development varies spontaneously (switches) in some moments of time which are random functionals of the previous trajectory. A special subclass of random processes with discrete component named switching processes were introduced by Anisimov in [5], [6], and [7].

Switching processes are two-component processes  $(x(t), \xi(t))$ ,  $t \geq 0$ , taking values in the space  $(X, R^r)$ , for which there exists a sequence of moments  $t_1 < t_2 < t_3 < \dots$  such that on each interval  $[t_k, t_{k+1})$ ,  $x(t) = x(t_k)$  and the behavior of the process  $\xi(t)$  depends on the value  $(x(t_k), \xi(t_k))$  only. The moments  $t_k$

are switching moments and  $x(t)$  is the discrete switching component. The component  $x(\cdot)$  usually corresponds to a random environment, a number of working servers or nodes (in queueing networks), etc.,  $\xi(t)$  can be the size of queue (or queues in the nodes of queueing networks), virtual waiting times, and process of lost calls, etc.

Switching processes are suitable for the analysis and asymptotic investigation of stochastic systems with “rare” or “fast” switchings. Various applications of switching systems can be found in Anisimov [7], Anisimov [11].

Anisimov in [19] considered the class of state-dependent queueing systems and networks with Markov or semi-Markov type switches and studied the results about the convergence of the vector-valued process  $\xi(t)$ , which corresponds to the number of calls in the system. The author proved the convergence to a solution of differential equation (Averaging principle) and to some diffusion process (Diffusion Approximation) in heavy traffic conditions for the case when the component  $x(t)$  is asymptotically ergodic. The method of investigation is based on limit theorems for so called switching processes [Anisimov [7] and [15]].

Another technique, called asymptotic merging of states, which allows us to study asymptotic characteristics for some classes of Markov systems with hierarchical arbitrary state space operating in different scales of time (slow and fast) was proposed by Anisimov [17]. Anisimov considered Markov systems of hierarchical structure functioning in different scales of time (slow and fast) and such that their local transition characteristics may be dependent on the current value of some other stochastic process (external random environment, stochastic failures, etc.). For systems of these types a new approach of decreasing dimension, approximate analytic modeling and estimating different reliability and efficiency characteristics is proposed.

Bobbio and Trivedi [21] introduced an approximation algorithm for systematically converting a stiff Markov chain into a nonstiff chain with smaller state space. After classifying the set of all states into fast and slow, the algorithm proceeds by further classifying fast states into fast recurrent subsets

and a fast transient subset. A separate analysis of each of these fast subsets is done and each fast recurrent subset is replaced by a single slow state while the fast transient subset is replaced by a probabilistic switch. After this reduction, the remaining small and nonstiff Markov chain is analyzed by a conventional technique.

Mostly asymptotic analysis for the time of loss of first customer and method of rare events was made for general queueing models of various types by Anisimov. Concerning the asymptotic analysis of retrial queueing models, only Falin [29] [30] [31], Stepanov [52] and Anisimov [18] performed some analyses.

Falin in [29] studied a retrial queueing system of the type  $M/M/C/\infty$  with absolutely insisting customers; the author studied the asymptotic behavior of the system's characteristics when the intensity of repetition becomes large. Later in [30], Falin studied the  $M/M/1$  retrial queueing system with no waiting space and derived some asymptotic results for heavy and light traffic. Falin in [31] studied the  $M/M/C/\infty$  retrial queueing system with loss (customers finding all servers busy are either queued or lost according to a Bernoulli switching rule). The author presented ergodic system occupancy results and ergodic server utilization under a high intensity of repetition for repeated calls.

Stepanov in [52] considered the  $M/M/s$  retrial queue with  $H_0 \leq 1$  and  $H_k = 1 (k \geq 1)$ . Stepanov considered the model in the case of extreme load and presented asymptotic formulation for the system. Characteristics such as the blocking probability, the mean number of busy servers, and the mean number of orbiting customers were obtained.

Anisimov in [18] studied transient and stable regimes in overloading retrial queueing systems. This approach is based on limit theorems of averaging principle and diffusion approximation types for so-called switching processes. Two models of retrial queueing systems of the types  $\bar{M}/\bar{G}/1$  with retrials (multidimensional Poisson input flow, one server with general service time, retrial system) and  $M/M/m$  with retrials ( $m$  servers with exponential service times) are considered in the case when the intensity of calls that reapply for the service tends to zero.

# Chapter 3

## PRELIMINARY RESULTS

In this chapter, an important notion of  $S$  – set (asymptotically connected set) is introduced. An exponential approximation for the first exit time from an  $S$  – set is introduced. A special class of hierarchical type  $S$  – sets, a “monotone structure”, is studied, and there is a part where the results obtained by Anisimov in [2] are studied. These results give us an analytical technique for the analysis and simulation of reliability characteristics of hierarchical Markov and semi-Markov models.

### 3.1 Results about asymptotic behavior of the first exit time from the fixed subset of states of SMP

Let  $x_{nk}, k \geq 0$  be a Markov process (MP) with finite state space  $X = \{1, 2, \dots, r\}$  depending on some parameter  $n$  and given by a matrix of one-step transition probabilities  $P_n = \|p_n(i, j)\|, i, j = \overline{1, r}$ . Let  $X_0$  be some fixed subset of  $X$ . Denote by

$$\nu_n(i) = \min\{k : k > 0, x_{nk} \notin X_0 \text{ given that } x_{n0} = i\}, i \in X_0, \quad (1)$$

the number of steps up to the time of the first exit from  $X_0$  starting from the state  $i \in X_0$ .

**Definition 3.1.1** *The subset  $X_0$  is called an S-set if for any  $i, j \in X_0$*

$\mathbf{P}\{ \text{there exists } k, k < \nu_n(i) \text{ such that } x_{nk} = j/x_{n0} = i \} \rightarrow 1 \text{ as } n \rightarrow \infty.$

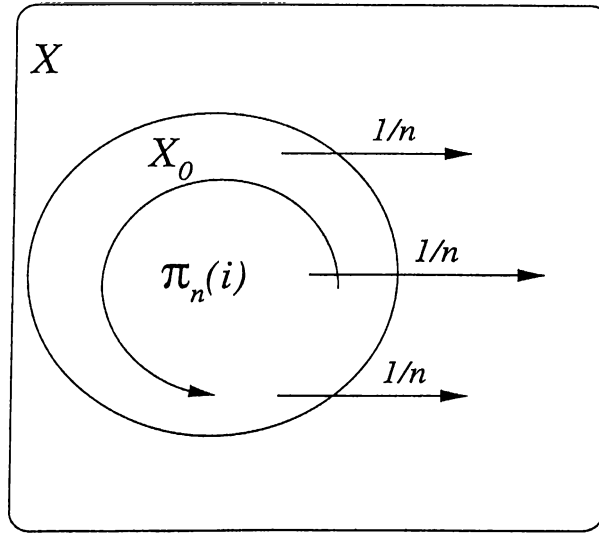


Figure 3.1: S-set

As can be seen from the Figure 3.1, the set  $X_0$  forms one essential class in limit ( $n \rightarrow \infty$ ) (meaning that the process will spend most of the time in subset  $X_0$  given that the initial state is  $i \in X_0$ ) and the probability that there exists  $k$  smaller than the number of steps to exit from the subset  $X_0$ , where  $k$  is the number of steps to go to state  $j \in X_0$  given that the process starts from the state  $i \in X_0$ , goes to one as  $n \rightarrow \infty$ .

Now let  $x_n(t)$  be a SMP with finite state space  $X = \{1, 2, \dots, r\}$  given by the embedded MP  $x_{nk}$  and by the family of sojourn times  $\{\tau_n(l), l = \overline{1, r}\}$  (suppose for simplicity that sojourn times do not depend on the next step). Denote by

$$\Omega_n(i) = \inf\{t : t > 0, x_n(t) \notin X_0 \text{ given that } x_n(0) = i\}$$

the time of the first exit from the subset  $X_0$  starting from the state  $i \in X_0$ .



Consider the limit behavior of the value  $\Omega_n(i)$ . Let us construct an auxiliary MP  $\tilde{x}_{nk}$  with state space  $X_0$  and matrix of transition probabilities  $\tilde{P}_n(X_0) = \|\tilde{p}_n(i, j)\|$ ,  $i, j \in X_0$  where

$$\tilde{p}_n(i, j) = p_n(i, j)p_n(i, X_0)^{-1}, \quad i, j \in X_0$$

and  $p_n(i, X_0)$  is defined as

$$p_n(i, X_0) = \sum_{l \in X_0} p_n(i, l).$$

Suppose that the set  $X_0$  forms an  $S$ -set. Denote by  $\tilde{\pi}_n(i)$ ,  $i \in X_0$  a stationary distribution for MP  $\tilde{x}_{nk}$  (which exists at least at large enough  $n$ ) and define

$$g_n(X_0) = \sum_{i \in X_0} \tilde{\pi}_n(i)(1 - p_n(i, X_0)). \quad (2)$$

**Theorem 3.1.1** *Let the set  $X_0$  form an  $S$ -set and there exist a normalizing factor  $\beta_n$  and functions  $a_i(\theta)$  ( $a_i(\pm 0) = 0$ ) such that as  $n \rightarrow \infty$*

$$g_n(X_0)^{-1}(1 - \mathbf{E} \exp\{-\beta_n \theta \tau_n(i)\}) \rightarrow a_i(\theta), \quad i \in X_0.$$

*Then for any initial state  $i_0 \in X_0$  it is true*

$$\lim_{n \rightarrow \infty} \mathbf{E} \exp\{-\beta_n \theta \Omega_n(i_0)\} = (1 + A(\theta))^{-1},$$

where

$$A(\theta) = \lim_{n \rightarrow \infty} \sum_{i \in X_0} \tilde{\pi}_n(i) a_i(\theta).$$

**Corollary 3.1.1** *In particular if the set  $X_0$  forms an  $S$ -set, then for any  $i_0 \in X_0$*

$$\lim_{n \rightarrow \infty} \mathbf{P}\{g_n(X_0) \nu_n(i_0) > t\} = \exp\{-t\}, \quad t > 0, \quad (3)$$

*which means that we have an exponential approximation for the number of steps in subset  $X_0$ .*

The proof can be found in Anisimov [2], [4], Anisimov et al. [9]. It is based on the asymptotic analysis of the matrix equation for the characteristic function

of the normed vector  $\{\beta_n \Omega_n(i), i \in X_0\}$  and uses the representation  $(I - \tilde{P}_n(X_0))^{-1} = g_n(X_0)^{-1} \tilde{\Pi}_n(X_0)$ , where  $I$  is the unit matrix, and  $\tilde{\Pi}_n(X_0) = \|\tilde{\pi}_n(i)(1 + o_{ij}(1))\|$ ,  $i, j \in X_0$ .

In these papers also an algorithm to check whether some subset forms an  $S$ -set or not is given. In papers of Anisimov [8] and Anisimov et al. [10] estimates of proximity for the rate of convergence in (3) are also given.

**Corollary 3.1.2** *Suppose that the process  $x_n(t), t \geq 0$  is a continuous time MP given by the embedded MP with matrix of transition probabilities  $P_n$  and by exit rates  $\lambda_n(i)$ ,  $i = \overline{1, r}$ , the set  $X_0$  forms an  $S$ -set,  $\min_i \lambda_n(i) \neq 0$  and  $\sum_{i \in X_0} \tilde{\pi}_n(i)/\lambda_n(i) \neq 0$ . In this case we can put  $\beta_n = g_n(X_0) \left( \sum_{i \in X_0} \tilde{\pi}_n(i)/\lambda_n(i) \right)^{-1}$  and the asymptotic distribution of the variable  $\beta_n \Omega_n(i)$  is exponential with parameter 1.*

We mention that in this case  $\beta_n$  is asymptotically equivalent to the value  $\sum_{i \in X_0} \tilde{\rho}_n(i) \sum_{k \notin X_0} \lambda_n(i, k)$ , where  $\tilde{\rho}_n(i), i \in X_0$  is the stationary distribution of the auxiliary continuous time MP with state space  $X_0$  and transition rates  $\lambda_n(i, j), i, j \in X_0, i \neq j$ .

These results show that to find a parameter in exponential approximation of exit time from the subset it is enough to estimate the main order of stationary probabilities  $\tilde{\pi}_n(i)$ ,  $i \in X_0$ .

If we denote by  $Y_n(t)$  the number of lost calls on the interval  $[0, t]$ . It can be said that the process  $Y_n(\beta_n^{-1}t)$  weakly converges in the sense of convergence of finite dimensional distribution to ordinary Poisson Process with some parameter. Asymptotic analysis of flows of rare events switched by some random environment is provided by Anisimov in [20]. The environment can be nonhomogeneous in time. In case when the environment satisfies an asymptotically mixing condition, an approximation by nonhomogeneous Poisson flows is proved. In general, it can be said that flows of rare events in systems with mixing can be approximated by Poisson process with average integral intensity

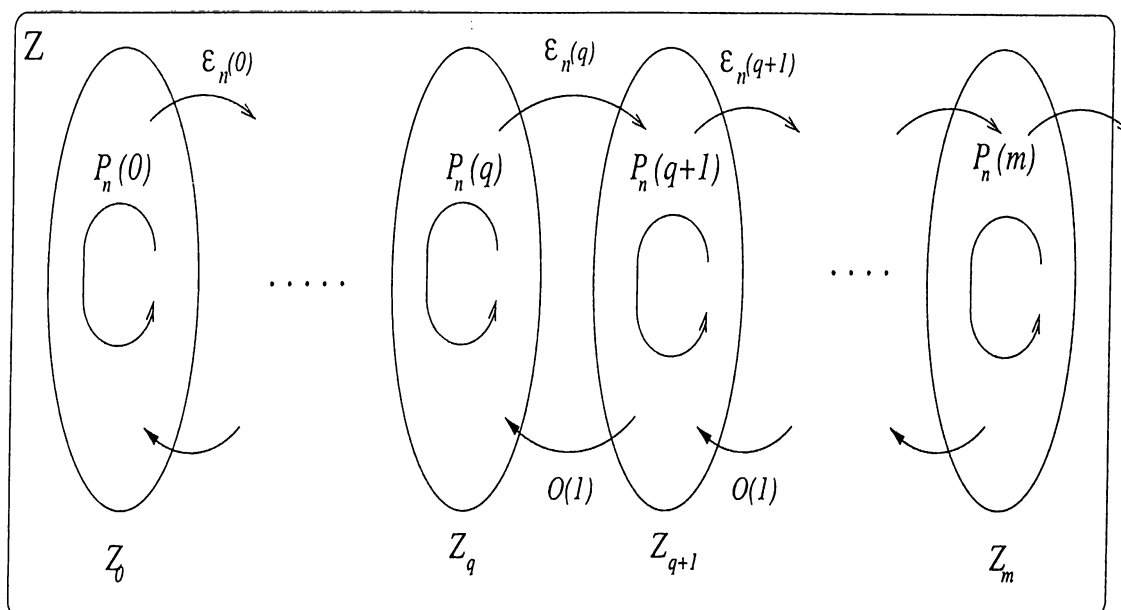


Figure 3.2: Monotone Structure

Let us consider an important class of  $MP$  with special monotone structure of its state space introduced in Anisimov et al.[9]. In this case it is possible to write an explicit formulas for main parts of stationary probabilities. Models of this type appear at the asymptotic analysis of wide classes of queueing systems with fast service.

Let  $x_{nk}, k \geq 0$  be some  $MP$  with finite state space  $Z = \{(l, q)\}$  and one-step transition probabilities  $p_n((i, s), (j, q))$ . Suppose that its state space can be represented in the form  $Z = \cup_{s=0}^{m+1} (Z_s, s)$ .

**Definition 3.1.2** *The subset of states  $Z = \{(i, s), i \in Z_s, s = \overline{0, m}\}$  is called a monotone structure if the following asymptotic relations hold:*

1.  $p_n((i, s), (j, s+1)) = \epsilon_n(s) a_{ij}(s) (1 + o(1)), i \in Z_s, j \in Z_{s+1}$ , where  $\epsilon_n(s) \rightarrow 0, s = \overline{0, m}$ ;
2.  $p_n((i, s), (j, s+k)) = 0, i \in Z_s, j \in Z_{s+k}, s = \overline{0, m-2}, k > 1$ ;
3.  $p_n((i, s), (j, s)) = p_{ij}(s) (1 + o(1)), i, j \in Z_s, s = \overline{0, m}$ ,

where the matrix  $I - P(s)$  is invertible for each  $s = \overline{1, m}$  and  $P(0)$  is an irreducible matrix with stationary distribution  $\pi_i$ ,  $i \in Z_0$  (here  $P(s) = \|p_{ij}(s)\|$ ,  $i, j \in Z_s$ ).

We call the subset of states  $Z_q = \{(i, q), i \in Z_q\}$  a  $q$ -level.

Figure 3.2 illustrates the general form of a monotone structure. Monotone structures, in a sense, can be interpreted as special type hierarchical  $S$ -sets (i.e, the subset  $Z$  forms an  $S$ -set). Condition 1 in the definition of the monotone structure ensures that the transition probabilities from any  $Z_q$  to  $Z_{q+1}$  for  $q = \overline{0, m}$  goes to zero in limit ( $n \rightarrow \infty$ ). Condition 2 ensures the monotonicity of the structure (i.e, there can be only transitions from  $Z_q$  to  $Z_{q+1}$  for  $q = \overline{0, m}$ , no transitions are possible from  $Z_q$  to  $Z_{q+s}$  for  $s > 1$ ). Condition 3 states that transition within any given level are possible and  $I - P(s)$  is invertible for each  $s = \overline{1, m}$  and  $P(0)$  is an irreducible matrix with stationary distribution where  $P(s) = \|p_{ij}(s)\|$ ,  $i, j \in Z_s$ . Transitions from any  $Z_q$  to  $Z_{q-1}$  are possible and probability of these transitions are of  $O(1)$  (see Figure 3.2 these transitions should exist, otherwise the matrix  $I - P(s)$  would be singular).

Let  $\bar{\pi}_n(s) = (\pi_n(i, s), i \in Z_s)$ ,  $s = \overline{0, m}$ ,  $\bar{\pi} = (\pi_i, i \in Z_s)$  and  $\bar{b} = (b_i, i = \overline{1, r})$  be row-vectors, where  $\pi_n(i, s)$  be the stationary probability of the state  $(i, s)$  for the  $MP$  with state space  $Z$  and matrix of transition probabilities

$$\tilde{P}_n(Z) = \|p_n((i, s), (j, q))p_n((i, s), Z)^{-1}\|, (i, s), (j, q) \in Z, \quad \cdot$$

where  $p_n((i, s), Z) = \sum_{(l, g) \in Z} p_n((i, s), (l, g))$ , and  $b_i = \sum_{k \in Z_{m+1}} a_{ik}(m)$ .

**Theorem 3.1.2** *If the state space  $Z = \{(i, s), i \in Z_s, s = \overline{0, m}\}$  forms a monotone structure then it also forms an  $S$ -set and for all  $q = \overline{1, m}$  the following representation holds:*

$$\bar{\pi}_n(q) = \bar{\pi} \left( \prod_{j=0}^{q-1} A(j)(I - P(j+1))^{-1} \epsilon_n(j) \right) (1 + o(1)), \quad (4)$$

and

$$g_n(Z) = \bar{\pi} \left( \prod_{j=0}^{m-1} A(j)(I - P(j+1))^{-1} \epsilon_n(j) \right) \epsilon_n(m) \bar{b}^* (1 + o(1)),$$

where  $A(s) = \|a_{ij}(s)\|$ ,  $i, j \in Z_s$ , and  $\bar{b}^*$  is the transposed vector to  $\bar{b}$ .

The main idea of the Theorem 3.1.2 is the following: In order to find the parameter of exponential approximation for the time of exit from a fixed subset of states, we need to find the stationary distribution of exit from  $Z$ ,  $(g_n(Z))$ . To be able to find  $g_n(Z)$  we need to know the stationary distribution of the states  $(\bar{\pi}_n(q), q = \overline{0, m})$ . Instead of solving the set of linear equations to find all of the stationary distributions, we just calculate the stationary distributions for the states in  $Z_0$  and multiply by the corresponding matrices to obtain  $g_n(Z)$ . Note that  $Z_0$  forms one essential class in limit (i.e, the process will spend most of the time in  $Z_0$ )

The proof of this result is made recursively to the order of the monotone structure. The main problem is in estimation of the stationary probabilities. It can be shown that

$$\pi_n(i, q) = O\left(\prod_{s=0}^{q-1} \varepsilon_n(s)\right), \quad i = \overline{1, r}, \quad q > 0, \quad \text{as } n \rightarrow \infty.$$

Then from the matrix equation

$$\bar{\pi}_n(q) = \bar{\pi}_n(q)P_n(q) + \bar{\pi}_n(q-1)\varepsilon_n(q-1)A(q-1) + O\left(\prod_{s=0}^q \varepsilon_n(s)\right),$$

where  $P_n(q) = \|p_n((i, q), (j, q))\|$ ,  $i, j \in X_q$ , we obtain

$$\bar{\pi}_n(q) = \pi_n(q-1)A(q-1)(I - P_n(q))^{-1}\varepsilon_n(q-1)(1 + o(1)),$$

and this implies (4). The expression for  $g_n(Z)$  follows from (2).

These results allow to study the asymptotic behavior of the time of first loss of a call for wide classes of queueing systems and networks with finite number of states and fast service or low loading (see Anisimov et al. [9], Anisimov [16]).

We note that as it follows from Theorem 3.1.1 the asymptotic behavior of a sojourn time in  $S$ -set does not depend on the initial state. This gives possibility to study models of asymptotic aggregation of state space (see Anisimov [3], Anisimov et al. [9]).

### 3.2 $M/M/s/m$ queueing system

In this section we consider an example which illustrates the use of the method of monotone structure. We will study the time of first customer loss in this system and will derive the parameter of exponential distribution using the method described in the previous section of this chapter.

Consider a Markov queueing system where customers arrive to the system with rate  $\lambda$  and there are  $s$  independent identical servers and the intensity of service is  $\mu_n$  for each server. The system has also  $m$  waiting places. The customer entering the system occupies the server or joins the queue. If all servers and all waiting positions are busy, the customer is lost.

Suppose that the service is asymptotically fast in the sense that

$$\mu_n = n\mu, \text{ where } n \rightarrow \infty$$

Let  $\Omega_n(q)$  be the time of the first loss of a customer if the initial conditions are  $Q_n(0) = q$  where  $Q_n(t)$  is the number of customers in the system at time  $t$ . We study the asymptotic behavior of  $\Omega_n(q)$  as  $n \rightarrow \infty$ .

$$\Omega_n(q) = \min\{t : t > 0, Q_n(t) > s + m\}$$

It is easy to see that the process  $Q_n(t)$  forms a homogeneous MP in continuous time and state space for the process is in the form  $I = \{(q), q = \overline{0, s + m}\}$

Then as  $q \leq s$

$$p_n(q, q + 1) = \frac{1}{n} \frac{\lambda}{q\mu}$$

otherwise, as  $q > s$

$$p_n(q, q + 1) = \frac{1}{n} \frac{\lambda}{s\mu}$$

It is also easy to see that the state space  $Z$  forms a monotone structure in which 0-level is the subset  $Z_0 = I_0 \cup I_1$ ,  $q$ -level is the subset  $Z_q = I_{q+1}$ ,  $0 < q \leq s + m$  and  $\varepsilon_n(s + m) = \frac{1}{n} \frac{\lambda}{s\mu}$

Let  $\pi_n(q)$  be the stationary distribution for the embedded MP.

$$\pi_n(q) = n^{-q+1} \pi_n(1) \frac{\lambda^{q-1}}{(q-1)! \mu^{q-1}} (1 + o(1)), \text{ if } q \leq s$$

and

$$\pi_n(q) = n^{-q+1} \pi_n(1) \frac{\lambda^{q-1}}{s! s^{q-s} \mu^{q-1}} (1 + o(1)), \text{ if } q > s$$

Applying the matrix relation of theorem 3.1.2, we obtain

$$g_n(Z) = n^{-s-m} \frac{\lambda^{s+m}}{2s! s^m \mu^{s+m}} (1 + o(1))$$

$g_n(Z)$  can be rewritten as

$$g_n(Z) = \frac{1}{n^{s+m}} G (1 + o(1))$$

where  $G$  is

$$G = \frac{\lambda^{s+m}}{2s! s^m \mu^{s+m}}$$

From Theorem 3.1.1 it follows

$$\frac{1}{n^{s+m}} G \Omega_n(i) \sim \hat{M} \eta_1 \quad \text{for any } i \in Z$$

where  $\hat{M} = \sum_{i \in Z_0} \pi_i \mathbf{E} \xi_i$ ,  $\eta_1$  is exponential distribution with parameter 1 and  $\xi_i$  is the sojourn time in state  $i$ . Rearranging the terms we get

$$\frac{1}{n^{s+m}} \Omega_n(i) \sim \frac{\hat{M}}{G} \eta_1$$

From above relations, by setting  $\beta_n = n^{-s-m}$ , we obtain the parameter of exponential distribution as

$$\Lambda \sim \frac{G}{\hat{M}}$$

**Theorem 3.2.1** *For the system  $M/M/s/m$  under the assumption of fast service the distribution of the variable  $n^{-s-m} \Omega_n(q)$  weakly converges for any  $0 \leq q \leq s+m$  to the exponential distribution with parameter*

$$\lim_{n \rightarrow \infty} \mathbf{P}\{n^{-s-m} \Omega_n(q) \geq t\} = \exp\{-At\}, \quad t > 0,$$

where

$$A = \frac{\lambda \rho^{s+m}}{s! s^m}$$

where  $\rho = \lambda/\mu$  and the flow of lost calls weakly converges to a Poisson one with parameter  $A$ .

## Chapter 4

# SINGLE-SERVER RETRIAL QUEUEING MODELS

### 4.1 $M/M/1/m$ system with retrials

Consider a single server queueing system with  $m$  waiting places in which customers arrive in a Poisson process with rate  $\lambda$ . These customers are identified as primary calls. If the server is free at the time of a primary call arrival, the arriving call begins to be served immediately and leaves the system after service completion. Otherwise, if the server is busy, the arriving customer becomes a source of repeated calls (a customer in retrial queue, a customer in orbit, a customer in pool, etc.). The pool of sources of repeated calls may be viewed as a sort of queue which we call retrial queue. Every such source produces a Poisson process of repeated calls with intensity of  $\nu$ . If an incoming repeated call finds the server free, it is served and leaves the system after service. On the other hand, if an arriving customer finds server and all of the waiting positions occupied, the customer will be lost. This system can be represented as an  $M/M/1/m$  system with retrials.

This chapter deals with the time of loss of first customer in an  $M/M/1/m$  queueing system with retrials under some different assumptions. The method



of analysis is based on the results about asymptotic behavior of the first exit time from the fixed subset of states of a *SMP* which forms an *S-set*, which we give in previous chapter. First we consider asymptotic behavior of the system under the assumption of fast service and then we consider the system under the assumptions of both fast service and fast retrials. Finally, we consider the system operating in Markov environment and under the assumption of fast service. We derive the expression for the parameter of exponential distribution for the time of loss of first customer for these models.

We assume the service time distribution is exponential with parameter  $\mu_n = n\mu$  (fast service as  $n \rightarrow \infty$ ) for both primary calls and repeated calls. Also we assume that the input flow of primary calls, intervals between repetitions, and service times are mutually independent.

The queueing process evolves in the following manner. Suppose that the  $(i - 1)$ th call completes its service at epoch  $\eta_{i-1}$  (the calls are numbered in the order of service) and the server becomes free. Even if there are some customers in the system who want to get service they cannot occupy the server immediately. Therefore the next,  $i$ th, call enters service only after some time interval  $R_i$  during which the server is free while there may be waiting customers. If the number of sources (number of customers in the queue) of repeated calls at the time  $\eta_{i-1}$  is equal to  $q$ , then the random variable  $R_i$  has an exponential distribution with parameter  $\lambda + q\nu$ . The  $i$ th call is a primary call with probability  $\frac{\lambda}{\lambda + q\nu}$  and it is a repeated call with probability  $\frac{q\nu}{\lambda + q\nu}$ . At epoch  $\xi_i = \eta_{i-1} + R_i$  the  $i$ th call's service starts and continues during a time  $S_i$  (service time of the  $i$ th call). All primary calls arriving during the service time form sources of repeated calls (i.e. join the retrial queue). Then, at epoch  $\eta_i = \xi_i + S_i$  the  $i$ th call completes service and the server becomes free again.

Let  $Q_n(t)$  be the number of sources of repeated calls (which may be viewed as a sort of retrial queue) at time  $t$  and  $\delta_n(t) = j, j = 0, 1$  denote the state of service at time  $t$  ( $\delta_n(t) = 1$  if in the moment  $t$   $i$ -th server is occupied and  $\delta_n(t) = 0$  otherwise). The process  $(\delta_n(t), Q_n(t))$  is Markov process and describes the number of customers in the system and is the simplest and simultaneously

the most important process associated with the above queueing system and the state space for the process is  $S = \{0, 1\} \times Z_+$ , where  $Z_+$  is the set of nonnegative integers.

Let  $\lambda, \nu_n$  and  $\mu_n$  be given and  $\lambda$  be the input rate,  $\nu_n$  be the rate for retrials for waiting customers and  $\mu_n$  be the service intensity where  $n$  is a scaling factor ( $n \rightarrow \infty$ ). We will consider the following cases:

**Case 1:**  $\nu_n = \nu$  (usual retrials) and  $\mu_n = n\mu$  (fast service) as  $n \rightarrow \infty$ .

**Case 2:**  $\nu_n = n\nu$  (fast retrials) and  $\mu_n = n\mu$  (fast service) as  $n \rightarrow \infty$ .

Denote by  $Q_n(t)$  the number of waiting calls (in the retrial queue) at time  $t$  also let  $Y_n(t)$  be the number of lost calls on the interval  $[0, t]$ .

**Theorem 4.1.1** *For the system described above (case 1), under the assumption of fast service, independently of the initial state, the distribution of the normalized random variable  $n^{-m-1}\Omega_n(j, q)$  weakly converges to an exponentially distributed random variable*

$$\lim_{n \rightarrow \infty} \mathbf{P}\{n^{-m-1}\Omega_n(j, q) \geq t\} = \exp\{-At\}, \quad t > 0,$$

where

$$A = \frac{\lambda \rho^{m+1}}{m! \nu^m} \prod_{k=1}^m (\lambda + k\nu),$$

where  $\lambda$  is the input rate,  $\nu$  is the rate for retrial calls and  $n\mu$  is the service intensity.

If the rate of incoming customers, rate of service and the rate of retrials depend on the size of the queue, the parameter of exponential distribution becomes

$$A = \lambda(0) \frac{1}{m!} \prod_{k=0}^m \frac{\lambda(k)}{\mu(k)} \prod_{k=0}^{m-1} \frac{\lambda(k) + (k+1)\nu(k+1)}{\nu(k+1)},$$

where  $\lambda(q)$  is the input rate,  $\nu(q)$  is the rate for retrial calls and  $n\mu(q)$  is the service intensity if  $Q_n(t) = q$ .

Also, the process  $Y(n^{m+1}t)$  weakly converges in the sense of convergence of finite dimensional distribution to ordinary Poisson Process with parameter  $A$ .

**Proof:**

Let  $\Omega_n(j, q)$  be the time of first loss of a call given  $Q_n(0) = q$  and  $\delta_n(0) = j$ . The asymptotic behavior of  $\Omega_n(j, q)$  as  $n \rightarrow \infty$  is studied.

$$\Omega_n(q) = \min\{t : t > 0, Q_n(t) > m\}$$

Consider a multicomponent process  $z_n(t) = (\delta_n(t), Q_n(t))$  where the indicator is introduced for the states of the server:  $\delta_n(t) = 1$  if at time  $t$ ,  $i$ -th server is occupied and  $\delta_n(t) = 0$  otherwise. The process  $z_n(t)$  forms a homogeneous MP in continuous time and the state space for the process is in the form of

$$Z = \{(j, q), j = 0, 1, q = \overline{0, m}\}$$

If we denote by  $\widehat{Q}_n(t)$  the number of waiting calls in the system with infinite number of waiting places, then  $\Omega_n(j, q)$  is the time of exit of the process  $\widehat{z}_n(t) = (\delta_n(t), \widehat{Q}_n(t))$  from the subset  $z_n(t)$ .

The rates of transitions for the process  $z_n(t)$  can be calculated and it can be seen that the subset  $Z$  forms monotone structure (see Definition 3.1.2) where at each fixed  $q = 0, 1, \dots, m$  the subset  $Z_q = \{(j, q), j = 0, 1\}$  forms  $q$ -level.

The monotone structure for the model and corresponding transition probabilities are shown in Figure 4.1 where  $\alpha_q$ ,  $\beta_q$  and  $\varepsilon_n(q)$  are defined as

$$\alpha_q = \frac{q\nu}{\lambda + q\nu} \quad \beta_q = \frac{\lambda}{\lambda + q\nu} \quad \varepsilon_n(q) = \frac{1}{n} \frac{\lambda}{\mu}$$

In each state  $(j, q)$  the process  $z_n(t)$  spends an exponential time with parameter

$$\Lambda(j, q) = \begin{cases} \lambda + n\mu & \text{if } j = 1 \\ \lambda + q\nu & \text{if } j = 0 \end{cases}$$

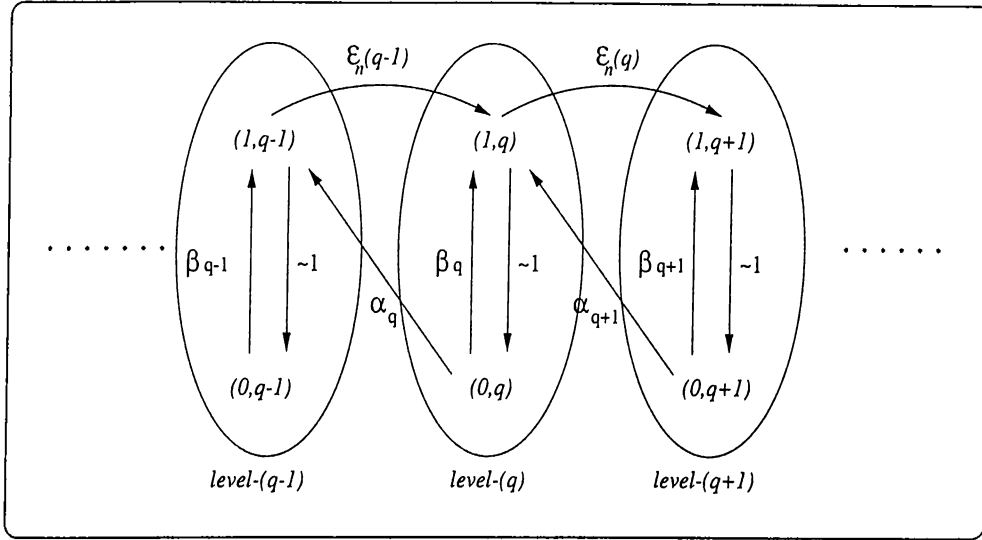


Figure 4.1: Monotone structure for single server model with assumption of fast service

The transition probabilities for the process are as follows:

$$\begin{aligned}
 p_n((1, q), (1, q+1)) &= \frac{1}{n} \frac{\lambda}{\mu} & p_n((0, q), (1, q-1)) &= \frac{q\nu}{\lambda + q\nu} \\
 p_n((0, q), (1, q)) &= \frac{\lambda}{\lambda + q\nu} & p_n((1, q), (0, q)) &= \frac{n\mu}{\lambda + n\mu} \rightarrow 1
 \end{aligned}$$

Now we can directly apply matrix relation of Theorem 3.1.2. Denote by  $\bar{\pi}_n(q) = (\pi_n(0, q), \pi_n(1, q))$  the stationary distribution of the embedded-Markov process for  $z_n(t)$  and let  $\pi_i, i = 0, 1$  ( $\bar{\pi} = (\pi_0, \pi_1)$ ) be the stationary distribution for the states at  $Z_0$  (level-0) of the monotone structure (such probabilities exist since  $Z_0$  (level-0) forms one essential class in limit).

The matrix relation of Theorem 3.1.2 is

$$\bar{\pi}_n(q) = \bar{\pi} \left( \prod_{j=0}^{q-1} A(j) (I - P(j+1))^{-1} \epsilon_n(j) \right) (1 + o(1)),$$

where matrices  $A(j)$  and  $P(j+1)$  are

$$A(j) = \begin{bmatrix} 0 & 0 \\ 0 & \lambda/\mu \end{bmatrix} \quad P(j+1) = \begin{bmatrix} 0 & \frac{\lambda}{\lambda + (j+1)\nu} \\ 1 & 0 \end{bmatrix}$$

The stationary distribution for the states of the embedded Markov process  $z_n(t)$  can be written as

$$\bar{\pi}_n(q) = \bar{\pi} \left( \prod_{j=0}^{q-1} \begin{bmatrix} 0 & 0 \\ 0 & \lambda/\mu \end{bmatrix} \left( I - \begin{bmatrix} 0 & \frac{\lambda}{\lambda+(j+1)\nu} \\ 1 & 0 \end{bmatrix} \right)^{-1} \frac{1}{n} \right) (1 + o(1)),$$

rearranging the terms, we obtain

$$\bar{\pi}_n(q) = \bar{\pi} \left( \prod_{j=0}^{q-1} \begin{bmatrix} 0 & 0 \\ 0 & \lambda/\mu \end{bmatrix} \frac{\lambda + (j+1)\nu}{(j+1)\nu} \begin{bmatrix} 1 & \frac{\lambda}{\lambda+(j+1)\nu} \\ 1 & 1 \end{bmatrix} \frac{1}{n} \right) (1 + o(1)).$$

Finally, the expression for  $\bar{\pi}_n(q)$  is obtained as follows

$$\bar{\pi}_n(q) = \pi_1 n^{-q} \frac{\rho^q}{q! \nu^q} \prod_{j=1}^q (\lambda + j\nu) (1 + o(1))$$

where  $\rho = \lambda/\mu$ , and note that

$$\bar{\pi}_n(q) = O \left( \prod_{s=0}^{q-1} \varepsilon_n(s) \right)$$

The expression for  $g_n(Z)$  can be obtained in the same way as applying the matrix relation of Theorem 3.1.2.

$$g_n(Z) = \pi_1 n^{-m-1} \frac{\rho^{m+1}}{m! \nu^m} \prod_{j=1}^m (\lambda + j\nu) (1 + o(1))$$

$g_n(Z)$  can be written as

$$g_n(Z) = \frac{1}{n^{m+1}} G$$

where  $G$  is

$$G = \pi_1 \frac{\rho^{m+1}}{m! \nu^m} \prod_{j=1}^m (\lambda + j\nu)$$

and it follows

$$\frac{1}{n^{m+1}} G \Omega(j, q) \sim \hat{M} \eta_1$$

where  $\hat{M} = \sum_{i \in Z_0} \pi_i \mathbf{E} \xi_i$ ,  $\eta_1$  is exponential distribution with parameter 1 and  $\xi_i$  is the sojourn time in state  $i$ . Rearranging the terms we get

$$\frac{1}{n^{m+1}} \Omega(j, q) \sim \frac{\hat{M}}{G} \eta_1$$

which means exponential approximation for the normalized variable  $n^{-m-1}\Omega(j, q)$ . From above relations, as setting the normalization constant  $\beta_n = n^{-m-1}$ , we obtain parameter of exponential distribution as

$$A \sim \frac{G}{\hat{M}}$$

which is

$$A = \frac{\lambda \rho^{m+1}}{m! \nu^m} \prod_{k=1}^m (\lambda + k\nu).$$

Now, we consider the system with single server and  $m$  waiting places where service and retrials are fast in the sense that  $\mu_n = n\mu$  and  $\nu_n = n\nu$  and will study the asymptotic behavior of the time of first lost customer as  $n \rightarrow \infty$ .

**Theorem 4.1.2** *For the system described above (case 2), under the assumption of fast service and fast retrials, independently of the initial state, the distribution of the normalized random variable  $n^{-m-1}\Omega_n(j, q)$  weakly converges to an exponentially distributed random variable*

$$\lim_{n \rightarrow \infty} \mathbf{P}\{n^{-m-1}\Omega_n(j, q) \geq t\} = \exp\{-At\}, \quad t > 0,$$

where

$$A = \lambda \left(\frac{\lambda}{\mu}\right)^{m+1}$$

where  $\lambda$  is the input rate,  $\nu$  is the rate for retrial calls, and  $n\mu$  is the service intensity.

If the rate of incoming customers, rate of service and the rate of retrials depend on the length of the queue, the parameter of exponential distribution becomes

$$A = \lambda(0) \prod_{k=0}^m \frac{\lambda(k)}{\mu(k)}$$

where  $\lambda(q)$  is the input rate,  $\nu(q)$  is the rate for retrial calls and  $n\mu(q)$  is the service intensity if  $Q_n(t) = q$ .

Also, the process  $Y(n^{m+1}t)$  weakly converges in the sense of convergence of finite dimensional distribution to ordinary Poisson Process with parameter  $A$ .

**Proof:**

The method of proof is similar to the previous one. Assume that the rate of service and the rate of retrial are fast in the sense that  $\mu_n = n\mu$  and  $\nu_n = n\nu$ .

$\Omega_n(j, q)$  is the time of first loss of a call given  $Q_n(0) = q$  and  $\delta_n(0) = j$ . We will consider the asymptotic behavior of  $\Omega_n(j, q)$  as  $n \rightarrow \infty$ .

Consider a multicomponent process  $z_n(t) = (\delta_n(t), Q_n(t))$  where indicator stands for the states of the server:  $\delta_n(t) = 1$  if in the time  $t$   $i$ -th server is occupied and  $\delta_n(t) = 0$  otherwise. The process  $z_n(t)$  forms a homogeneous MP in continuous time and the state space for the process is in the form of

$$Z = \{(j, q), j = 0, 1, q = \overline{0, m}\}$$

The subset  $Z$  forms monotone structure where at each fixed  $q = 0, 1, \dots, m$  the subset  $Z_q = \{(j, q), j = 0, 1\}$  forms  $q$ -level.

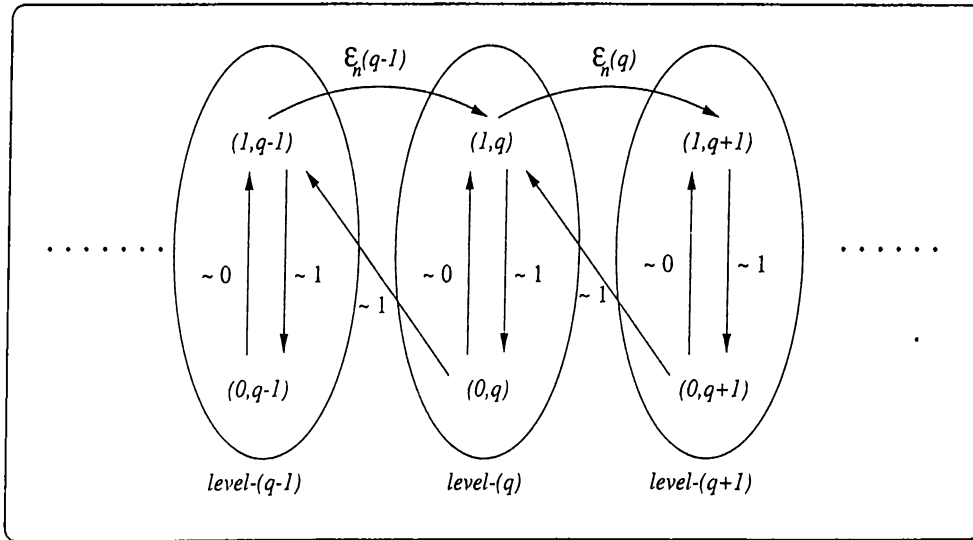


Figure 4.2: Monotone structure for the model with single server and assumptions of fast service and fast retrials

Figure 4.2 illustrates the monotone structure for the model and corresponding transition probabilities.

In each state  $(j, q)$  the process  $z_n(t)$  spends an exponential time with

parameter

$$\Lambda(j, q) = \begin{cases} \lambda + jn\mu & \text{if } j = 1 \\ \lambda + qn\nu & \text{if } j = 0 \end{cases}$$

The transition probabilities for the process are as follows:

$$\begin{aligned} p_n((1, q), (1, q + 1)) &= \frac{1}{n} \frac{\lambda}{\mu} & p_n((0, q), (1, q - 1)) &= \frac{qn\nu}{\lambda + qn\nu} \rightarrow 1 \\ p_n((0, q), (1, q)) &= \frac{\lambda}{\lambda + qn\nu} \rightarrow 0 & p_n((1, q), (0, q)) &= \frac{n\mu}{\lambda + n\mu} \rightarrow 1 \end{aligned}$$

Denote by  $\bar{\pi}_n(q) = (\pi_n(0, q), \pi_n(1, q))$  the stationary distribution of the embedded Markov process for  $z_n(t)$  and let  $\pi_i, i = 0, 1$  ( $\bar{\pi} = (\pi_0, \pi_1)$ ) be the stationary distribution for the process  $\delta_n(t)$ .

Applying the matrix relation of Theorem 3.1.2,

$$\bar{\pi}_n(q) = \bar{\pi} \left( \prod_{j=0}^{q-1} A(j) (I - P(j+1))^{-1} \varepsilon_n(j) \right) (1 + o(1)),$$

where matrices  $A(j)$  and  $P(j+1)$  are

$$A(j) = \begin{bmatrix} 0 & 0 \\ 0 & \frac{\lambda}{\mu} \end{bmatrix} \quad P(j+1) = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$$

The expression for  $\bar{\pi}_n(q)$  is obtained as

$$\bar{\pi}_n(q) = \pi_1 \frac{1}{n^q} \left( \frac{\lambda}{\mu} \right)^q (1 + o(1))$$

The expression for  $g_n(Z)$  follows from Theorem 3.1.2

$$g_n(Z) = \frac{1}{n^{m+1}} \pi_1 \left( \frac{\lambda}{\mu} \right)^{m+1} (1 + o(1))$$

Finally as setting  $\beta_n = n^{-m-1}$  we obtain the parameter of exponential distribution as

$$A = \lambda \left( \frac{\lambda}{\mu} \right)^{m+1}$$

The same result can be obtained as substituting  $\nu$  in the formula obtained for Theorem 4.1 with  $\nu_n = n\nu$  and analyze the case when  $n \rightarrow \infty$ . Also, note that the result obtained does not depend on the retrial rate  $\nu$ .



## 4.2 $M_U/M_U/1/m$ retrial queueing system operating in Markov environment

Consider a Markov retrial queueing system of the type  $M_U/M_U/1/m$  with retrials. The system with one server and  $m$  waiting places. Calls enter the system one at a time. On arrival, if the server is free the customer will receive service immediately; otherwise, if there are free waiting positions the customer will join the queue waiting for service. On the other hand, if an arriving customer finds server and all of the waiting positions occupied, the customer will be lost. Each waiting customer independently of others repeats its attempts for service after some random time. If at this time the server is free it takes the customer for service, if server is busy the call remains in the queue and repeats its attempts for service in the same way.

Suppose that the system is operating in a Markov environment  $x(t), t \geq 0$  with finite state space  $X = \{1, 2, \dots, r\}$  given by some initial state  $x_0$  and rates of transitions  $a_{ij}, i, j \in X, i \neq j$ .

Let  $\lambda(i, q), \nu(i, q), n\mu(i, q), i \in X, q = \overline{0, m}$  be given non-negative functions and  $\lambda(i, q)$  be the instantaneous input rate,  $\nu(i, q)$  be the rate for retrials for waiting customers and  $n\mu(i, q)$  be the service intensity given  $x(t) = i$  and  $Q_n(t) = q$  where  $n$  is a scaling factor ( $n \rightarrow \infty$ ). That means the service is fast. Denote by  $Q_n(t)$  the number of waiting calls (in the retrial queue) at time  $t$ . Denote also by  $Y_n(t)$  the number of lost calls on the interval  $[0, t]$ .

Now, we will study the asymptotic behavior of the time of loss of first customer and will derive the parameter of the exponential approximation for the time of loss of first customer using the same technique (see Definition 3.1.2)

Let  $\Omega_n(i, j, q)$  be the time of first loss of a call given  $Q_n(0) = q$  and  $\delta_n(0) = j$ . The asymptotic behavior of  $\Omega_n(i, j, q)$  as  $n \rightarrow \infty$  is studied.

Consider a multicomponent process  $z_n(t) = (x(t), \delta_n(t), Q_n(t))$  where indicator is introduced for the states of the server:  $\delta_n(t) = 1$  if in the moment

$t$   $i$ -th server is occupied and  $\delta_n(t) = 0$  otherwise. The process  $z_n(t)$  forms a homogeneous MP in continuous time and the state space for the process is in the form of

$$Z = \{(i, j, q), i \in X, j = 0, 1, q = \overline{0, m}\}$$

If we denote by  $\widehat{Q}_n(t)$  the number of waiting calls in the system with infinite number of waiting places, then  $\Omega_n(i, j, q)$  is the time of exit of the process  $z_n(t)$  from the subset  $\widehat{z}_n(t)$ .

The rates of transitions for the process  $z_n(t)$  can be calculated and it can be seen that the subset  $Z$  forms monotone structure where at each fixed  $q = 0, 1, \dots, m$  the subset  $Z_q = \{(i, j, q), i = \overline{1, r}, j = 0, 1\}$  forms  $q$ -level.

Monotone structure and corresponding transition probabilities for the model described above are shown in Figure 4.3.

In each state  $(i, j, q)$  the process  $z_n(t)$  spends an exponential time with parameter

$$\Lambda(i, j, q) = \begin{cases} \lambda(i, q) + n\mu(i, q) + a_{ii} & \text{if } j = 1 \\ \lambda(i, q) + q\nu(i, q) + a_{ii} & \text{if } j = 0 \end{cases}$$

where  $a_{ii} = \sum_{k \neq i} a_{ik}$ . The transition probabilities for the process are as follows:

$$p_n((i, 1, q), (i, 1, q+1)) = \frac{\lambda(i, q)}{\lambda(i, q) + a_{ii} + n\mu(i, q)} \rightarrow \frac{1}{n} \frac{\lambda(i, q)}{\mu(i, q)} \text{ for } i = \overline{1, r},$$

$$p_n((i, 0, q), (i, 1, q-1)) = \frac{q\nu(i, q)}{\lambda(i, q) + a_{ii} + q\nu(i, q)},$$

$$p_n((i, 0, q), (i, 1, q)) = \frac{\lambda(i, q)}{\lambda(i, q) + a_{ii} + q\nu(i, q)},$$

$$p_n((i, 0, q), (k, 0, q)) = \frac{a_{ik}}{\lambda(i, q) + a_{ii} + q\nu(i, q)}, \quad i \neq k,$$

$$p_n((i, 1, q), (i, 0, q)) = \frac{n\mu(i, q)}{\lambda(i, q) + a_{ii} + n\mu(i, q)} \rightarrow 1,$$

$$p_n((i, 1, q), (k, 1, q)) = \frac{a_{ik}}{\lambda(i, q) + a_{ii} + n\mu(i, q)} \rightarrow 0, \quad i \neq k,$$

Now we can directly apply matrix equation of Theorem 3.1.2. Denote by  $\bar{\pi}_n(j, q) = (\pi_n(i, j, q), i \in X, j = 0, 1, q = 0, 1, \dots, m)$  the stationary distribution

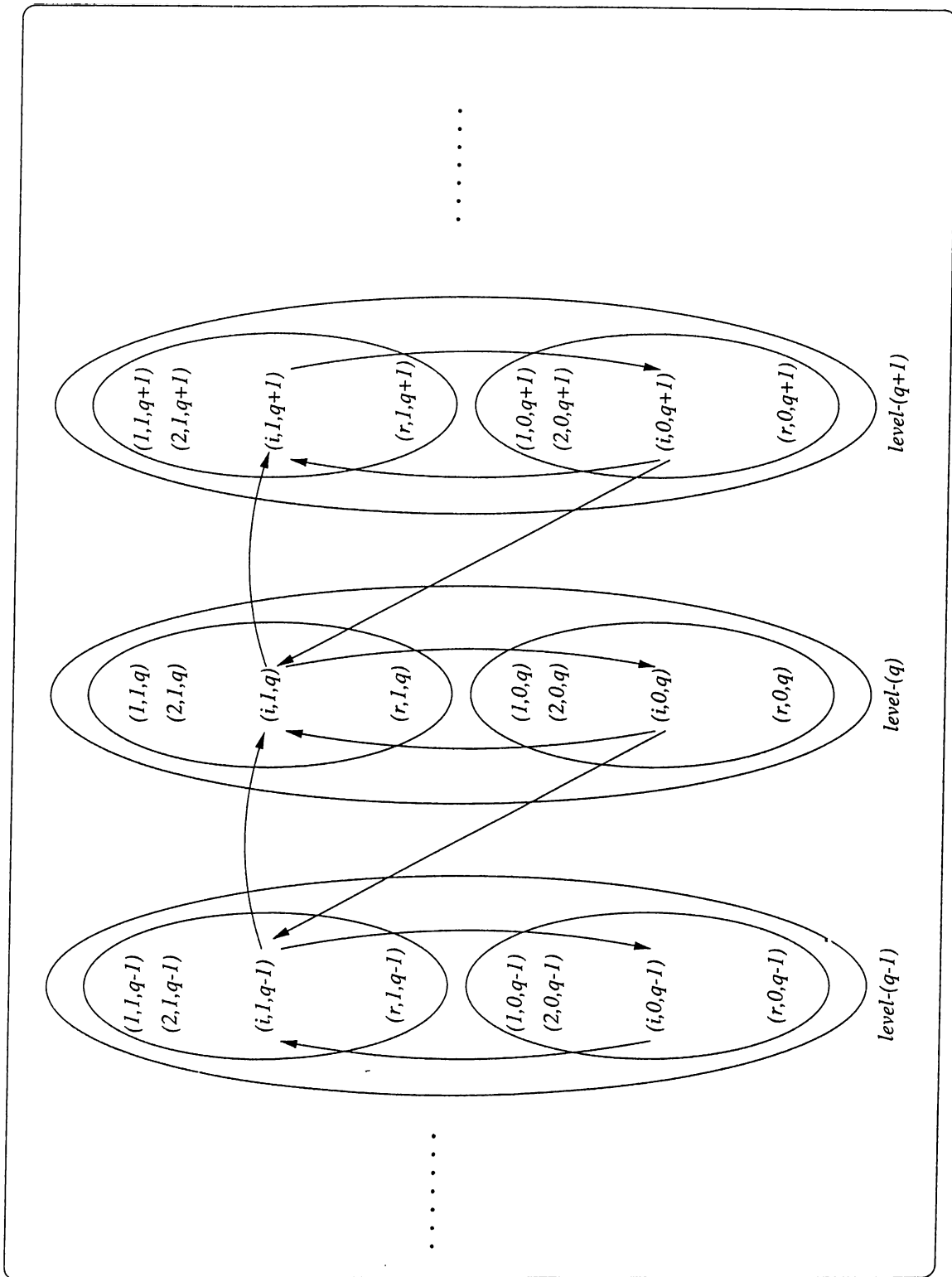


Figure 4.3: Monotone structure for single server system which operates in additional Markov environment and assumption of fast service

of the embedded Markov process for  $z_n(t)$  and let  $\pi_i, i = \overline{1, r}$  be the stationary distribution for the process  $x(t)$  and  $\bar{\pi}$  be the row vector  $(\pi_1, \dots, \pi_r)$ .

The matrix equation is

$$\bar{\pi}_n(q) = \bar{\pi} \left( \prod_{j=0}^{q-1} A(j)(I - P(j+1))^{-1} \varepsilon_n(j) \right) (1 + o(1)),$$

where the matrix  $A(j)$  is defined as

$$A(j) = \begin{bmatrix} 0 & 0 \\ 0 & G(j) \end{bmatrix}$$

and  $G(q)$  is a matrix with elements  $\lambda(i, q)/\mu(i, q)$  on diagonal and  $P(j+1)$  is defined as

$$P(j+1) = \begin{bmatrix} B(j+1) & \Lambda(j+1) \\ I & 0 \end{bmatrix}$$

where  $B(q)$  is defined as

$$B(q) = \left\| \frac{a_{ij}(1 - \delta_{ij})}{\lambda(i, q) + a_{ii} + q\nu(i, q)} \right\|, \quad i, j = \overline{1, r},$$

where  $\delta_{ij} = 0$  if  $i = j$  and  $\delta_{ij} = 1$  if  $i \neq j$  and  $\Lambda(q)$  is defined as

$$\Lambda(q) = \left\| \frac{\lambda(i, q)\delta_{ij}}{\lambda(i, q) + a_{ii} + q\nu(i, q)} \right\|, \quad i, j = \overline{1, r}, \quad q = \overline{1, m}$$

Define  $\bar{\pi}_n(q) = (\bar{\pi}_n(0, q), \bar{\pi}_n(1, q))$  where  $\bar{\pi}_n(0, q) = (\pi(i, 0, q), i = \overline{1, r}, q = \overline{0, m})$  and  $\bar{\pi}_n(1, q) = (\pi(i, 1, q), i = \overline{1, r}, q = \overline{0, m})$  are row vectors.

$$(I - P(q))^{-1} = \begin{bmatrix} (I - B(q) - \Lambda(q))^{-1} & (I - B(q) - \Lambda(q))^{-1} \Lambda(q) \\ (I - B(q) - \Lambda(q))^{-1} & (I - B(q) - \Lambda(q))^{-1} (I - B(q)) \end{bmatrix}.$$

and expression for the stationary distribution becomes

$$\bar{\pi}_n(q) = [\bar{\pi}(0, 0), \bar{\pi}(1, 0)] \left( \prod_{j=0}^{q-1} \begin{bmatrix} 0 & 0 \\ G(j)K(j+1) & G(j)K(j+1)(I - B(j+1)) \end{bmatrix} \right) \frac{1}{n^q} \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

where  $K(j) = (I - B(j) - \Lambda(j))^{-1}$  and following the equation for  $g_n(Z)$  of Theorem 3.1.2, we obtain

$$g_n(Z) = \frac{1}{n^{m+1}} \bar{\pi}(1, 0) \left( \prod_{j=0}^{m-1} G(j)K(j+1)(I - B(j+1)) \right) G(m)(1 + o(1)).$$

Since the level  $Z_0$  forms in limit one essential class, stationary distribution for each state in  $Z_0$  exist and satisfy the system of equations:

$$\bar{\pi}(0, 0) = \sum_{k \neq i} \pi(k, 0, 0) \frac{a_{ki}}{\lambda(k, 0) + a_{kk}} + \bar{\pi}(1, 0)$$

$$\bar{\pi}(1, 0) = \bar{\pi}(0, 0) \frac{\lambda(i, 0)}{\lambda(i, 0) + a_{ii}}, \quad i = \overline{1, r}$$

It can be easily shown that

$$\pi(i, 0, 0) = \frac{(\lambda(i, 0) + a_{ii})}{\sum_{k=1}^r \pi_k (2\lambda(k, 0) + a_{kk})} \pi_i, \quad i = \overline{1, r}$$

and

$$\pi(i, 1, 0) = \frac{\lambda(i, 0)}{\sum_{k=1}^r \pi_k (2\lambda(k, 0) + a_{kk})} \pi_i, \quad i = \overline{1, r}$$

Finally we obtain the parameter of exponential distribution as

$$A = \bar{\pi} \Lambda(0) G(0) (I - B(1) - \Lambda(1))^{-1} (I - B(1)) G(1) \dots \\ \dots G(m-1) (I - B(m) - \Lambda(m))^{-1} (I - B(m)) G(m) \bar{1},$$

**Theorem 4.2.1** *At our conditions for any initial state  $(i, j, q) \in Z$*

$$\lim_{n \rightarrow \infty} \mathbf{P}\{n^{-m-1} \Omega_n(i, j, q) \geq t\} = \exp\{-At\}, \quad t > 0,$$

where

$$A = \bar{\pi} \Lambda(0) G(0) (I - B(1) - \Lambda(1))^{-1} (I - B(1)) G(1) \dots \\ \dots G(m-1) (I - B(m) - \Lambda(m))^{-1} (I - B(m)) G(m) \bar{1},$$

and the process  $Y(n^{m+1}t)$  weakly converges in the sense of convergence of finite dimensional distribution to ordinary Poisson Process with parameter  $A$ .

## Chapter 5

# MULTIPLE-SERVER RETRIAL QUEUEING MODELS

Consider a group of  $s$  fully available servers in which a Poisson flow of calls with rate  $\lambda$  arrives and system consists of  $m$  waiting places for repeated calls. If an arriving primary call finds some server free, it immediately occupies a server and leaves the system after service. Otherwise, if all servers are engaged, an arriving primary call produces a source of repeated calls. Every such source after some delay produces repeated calls until after one or more attempts it finds a free server, in which case the call receives service and then leaves the system.

We assume that periods between successive retrials are exponentially distributed with parameter  $\nu_n$ , and service times are exponentially distributed with parameter  $\mu_n$ . As usual, we suppose that interarrival periods, retrial times and service times are mutually independent.

The functioning of the system can be described by means of a bivariate process  $(N_n(t), Q_n(t))$ , where  $N_n(t)$  is the number of busy servers and  $Q_n(t)$  is the number of calls in the retrial queue at time  $t$ . Under the above assumptions

the bivariate process  $(N_n(t), Q_n(t))$  is Markovian with the state space  $S = \{0, 1, \dots, s\} \times Z_+$

In this chapter, we study multiple-server retrial queueing system of type  $M/M/s/m$  and will derive the expression for the parameter of exponential distribution for the time of loss of first customer. We will consider two cases:

**Case 1:**  $\nu_n = \nu$  (usual retrials) and  $\mu_n = n\mu$  (fast service) as  $n \rightarrow \infty$ .

**Case 2:**  $\nu_n = n\nu$  (fast retrials) and  $\mu_n = n\mu$  (fast service) as  $n \rightarrow \infty$ .

The method of derivation is as follows: first we study the system with only 2 servers ( $s = 2$ ) and  $m$  waiting places, then the general result for  $s$  server case is obtained as generalizing the previous results.

## 5.1 $M/M/2/m$ system with retrials

Consider Markov retrial queueing system with two independent and identical servers and  $m$  waiting places. Customers arrive to the system one at a time and customers arrive according to Poisson process with rate  $\lambda$ . On arrival, if a customer finds one of the servers free, he will be served with an exponential rate of  $\mu_n$  immediately, otherwise he will join the special queue from where he will repeat, independently of other customers, his attempts for service after an exponential time with rate  $\nu_n$ . On the other hand, if an arriving customer finds server and all of the waiting positions occupied, the customer will be lost.

Let  $\lambda, \nu$  and  $\mu_n$  be given and  $\lambda$  be the input rate,  $\nu$  be the rate for retrials for waiting customers and  $\mu_n = n\mu$  be the service intensity for each server where  $n$  is a scaling factor ( $n \rightarrow \infty$ ). That means the service is fast. Denote by  $Q_n(t)$  the number of waiting calls (in the retrial queue) at time  $t$ .

**Theorem 5.1.1** *For the system described above (case 1), under the assumption of fast service, independently of the initial state, the distribution of the*

normalized random variable  $n^{-m-2}\Omega_n(j, q)$  converges weakly to an exponentially distributed random variable

$$\lim_{n \rightarrow \infty} \mathbf{P}\{n^{-m-2}\Omega_n(j, q) \geq t\} = \exp\{-At\}, \quad t > 0,$$

where

$$A = \lambda \left(\frac{\lambda}{2\mu}\right)^{m+2}$$

where  $\lambda$  is the input rate,  $\nu$  is the rate for retrial calls, and  $\mu_n = n\mu$  is the service intensity.

**Proof:**

Let  $\Omega_n(j, q)$  be the time of first loss of a call given  $Q_n(0) = q$  and  $N_n(0) = j$  in a system where the rate of service is fast in the sense that  $\mu_n = n\mu$ . We study the asymptotic behavior of  $\Omega_n(j, q)$  as  $n \rightarrow \infty$ .

Consider a multicomponent process  $(N_n(t), Q_n(t))$  with state space  $S = \{0, 1, 2\} \times Z_+$ . The process  $(N_n(t), Q_n(t))$  forms a homogeneous MP in continuous time and the state space for the process is in the form of

$$Z = \{(j, q), j = 0, 1, 2, q = \overline{0, m}\}$$

The rates of transitions for the process  $(N_n(t), Q_n(t))$  can be calculated and it can be seen that the subset  $Z$  forms monotone structure where at each fixed  $q = 0, 1, \dots, m$  the subset  $Z_q = \{(j, q), j = 0, 1, 2\}$  forms  $q$ -level.

Figure 5.1 shows the monotone structure and corresponding probabilities where  $\alpha_q, \beta_q, a_q, b_q$  and  $\varepsilon_n(q)$  are defined as

$$\alpha_q = \frac{q\nu}{\lambda + q\nu} \qquad \beta_q = \frac{\lambda}{\lambda + q\nu}$$

$$a_q = \frac{1}{n} \frac{q\nu}{\mu} \rightarrow 0 \qquad b_q = \frac{1}{n} \frac{\lambda}{\mu} \rightarrow 0 \qquad \varepsilon_n(q) = \frac{1}{n} \frac{\lambda}{2\mu}$$

and note that  $a_q$  and  $b_q$  are in the order of  $\varepsilon_n(q)$ .



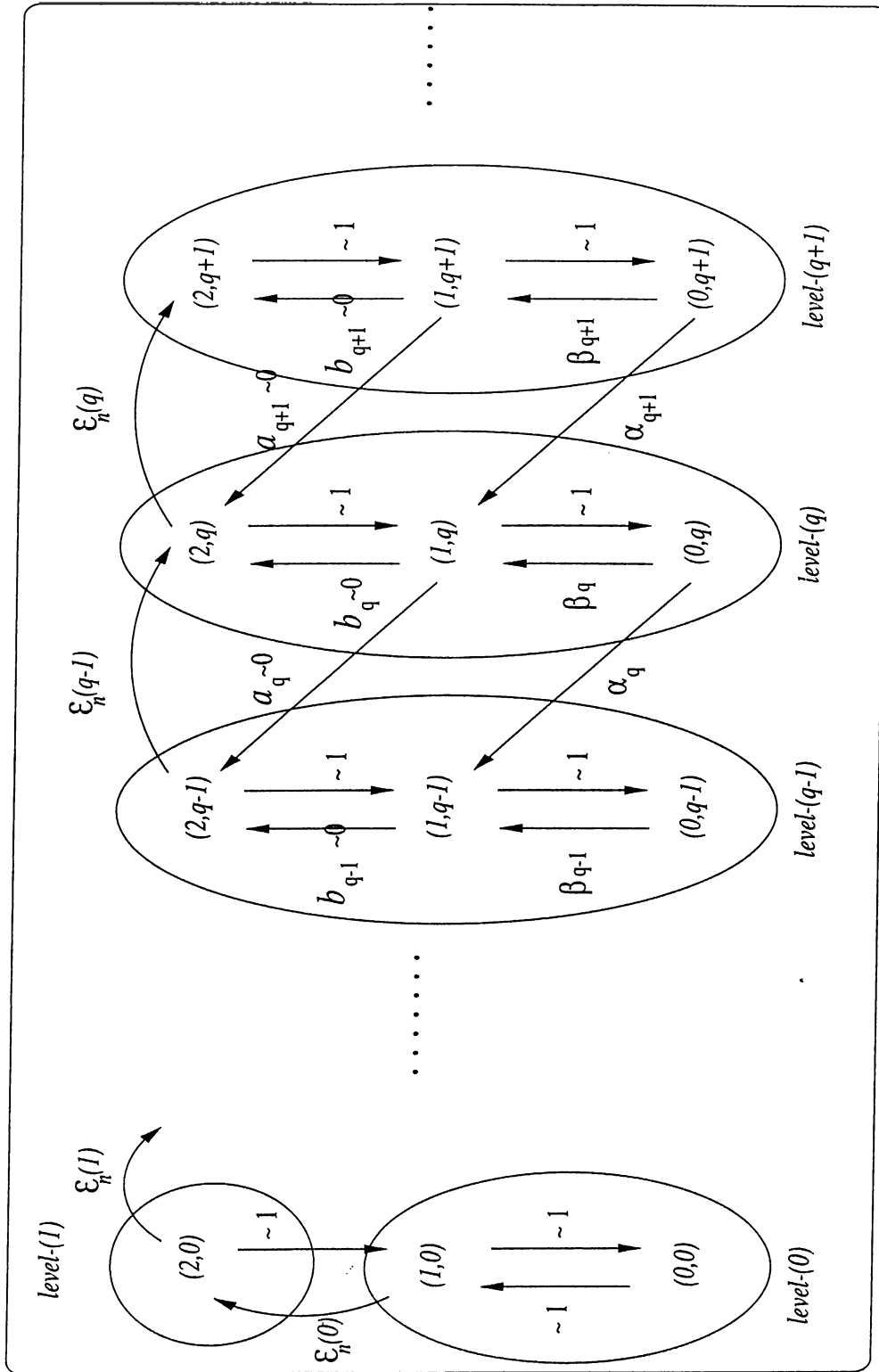


Figure 5.1: Monotone structure for the model with two servers and assumption of fast service

In each state  $(j, q)$  the process spends an exponential time with parameter

$$\Lambda(j, q) = \begin{cases} \lambda + 2n\mu & \text{if } j = 2 \\ \lambda + n\mu + q\nu & \text{if } j = 1 \\ \lambda + q\nu & \text{if } j = 0 \end{cases}$$

Denote by  $\bar{\pi}_n(q) = (\pi_n(0, q), \pi_n(1, q), \pi_n(2, q))$  for  $q = 2, 3, \dots, m + 2$  the stationary distribution of the embedded Markov process for  $(N_n(t), Q_n(t))$  and let  $\pi_i, i = 0, 1, 2$  ( $\bar{\pi} = (\pi_0, \pi_1, \pi_2(n))$ ) be the stationary distribution for the states in  $Z_0$  and  $Z_1$  (see Figure 5.1) where components  $\pi_0$  and  $\pi_1$  belong to the states in  $Z_0$  and  $\pi_2(n)$  belongs to the state in  $Z_1$ . Note that  $Z_0$  in limit forms one essential class. The method of study will be as follows: First, we will obtain expression for  $g_n(Z)$  which will be a function of  $\pi_2(n)$ , then we will use  $\pi_0$  and  $\pi_1$  to recalculate  $\pi_2(n)$  and obtain final expression for  $g_n(Z)$ .

The expression for  $g_n(Z)$  can be obtained directly as applying matrix relation of Theorem 3.1.2

$$g_n(Z) = \pi_2(n) \frac{1}{n^{m+1}} \left( \frac{\lambda}{2\mu} \right)^{m+1} (1 + o(1))$$

hence, as substituting the  $\pi_2(n) = \frac{1}{2} \frac{1}{n} \frac{\lambda}{2\mu}$  in the relation above, we obtain

$$g_n(Z) = \frac{1}{n^{m+2}} \frac{1}{2} \left( \frac{\lambda}{2\mu} \right)^{m+2} (1 + o(1))$$

and from the expression for  $g_n(Z)$  we obtain, as setting  $\beta_n = n^{-m-2}$ , the parameter of exponential distribution as

$$A = \lambda \left( \frac{\lambda}{2\mu} \right)^{m+2}$$

Now, we will study the same system described above with assumption of fast service and fast retrials which is  $\mu_n = n\mu$  and  $\nu_n = n\nu$  and will study the asymptotic behavior of the time of first lost customer as  $n \rightarrow \infty$ .

**Theorem 5.1.2** *For the system described above (case 2), under the assumption of fast service and fast retrials, independently of the initial state, the distribution of the normalized random variable  $n^{-m-2}\Omega_n(j, q)$  converges weakly*

to an exponentially distributed random variable

$$\lim_{n \rightarrow \infty} \mathbf{P}\{n^{-m-2}\Omega_n(j, q) \geq t\} = \exp\{-At\}, \quad t > 0,$$

where

$$A = \lambda \left(\frac{\lambda}{2\mu}\right)^{m+2}$$

where  $\lambda$  is the input rate,  $n\nu$  is the rate for retrial calls, and  $n\mu$  is the service intensity.

### Proof

Let  $\Omega_n(j, q)$  be the time of first loss of a call given  $Q_n(0) = q$  and  $N_n(0) = j$  in a system where the rate of service and the rate of retrials are fast in the sense that  $\mu_n = n\mu$  and  $\nu_n = n\nu$ . The asymptotic behavior of  $\Omega_n(j, q)$  as  $n \rightarrow \infty$  is studied.

Consider a multicomponent process  $(N_n(t), Q_n(t))$  with state space  $S = \{0, 1, 2\} \times Z_+$ . The process  $(N_n(t), Q_n(t))$  forms a homogeneous MP in continuous time and the state space for the process is in the form of

$$Z = \{(j, q), j = 0, 1, 2, q = \overline{0, m}\}$$

The rates of transitions for the process  $(N_n(t), Q_n(t))$  can be calculated and it can be seen that the subset  $Z$  forms monotone structure where at each fixed  $q = 0, 1, \dots, m$  the subset  $Z_q = \{(j, q), j = 0, 1, 2\}$  forms  $q$ -level.

Monotone structure for the system can be seen in Figure 5.2 and  $\varepsilon_n(q)$  is defined as

$$\varepsilon_n(q) = \frac{1}{n} \frac{\lambda}{2\mu}$$

In each state  $(j, q)$  the process spends an exponential time with parameter

$$\Lambda(j, q) = \begin{cases} \lambda + 2n\mu & \text{if } j = 2 \\ \lambda + n\mu + qn\nu & \text{if } j = 1 \\ \lambda + qn\nu & \text{if } j = 0 \end{cases}$$

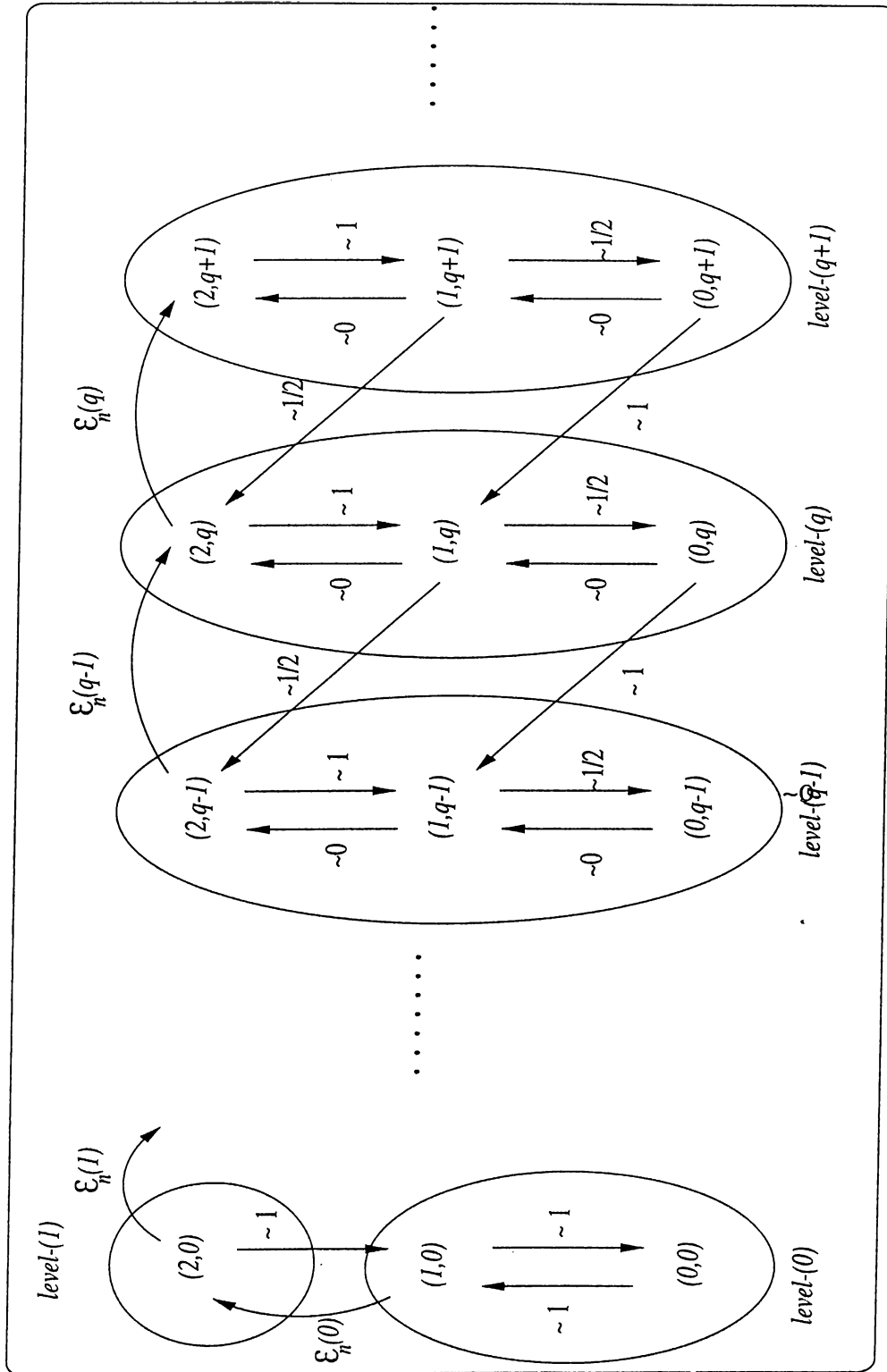


Figure 5.2: Monotone structure for the model with two servers and assumptions of fast service and fast retrials

Denote by  $\bar{\pi}_n(q) = (\pi_n(0, q), \pi_n(1, q), \pi_n(2, q))$  for  $q = 2, 3, \dots, m + 2$ , the stationary distribution of the embedded Markov process for  $(N_n(t), Q_n(t))$  and let  $\pi_i, i = 0, 1, 2$  ( $\bar{\pi} = (\pi_0, \pi_1, \pi_2(n))$ ) be the stationary distribution for the states in  $Z_0$  and  $Z_1$  (see Figure 5.2) where components  $\pi_0$  and  $\pi_1$  belong to the states in  $Z_0$  and  $\pi_2(n)$  belongs to the state in  $Z_1$ . Note that  $Z_0$  in limit forms one essential class. The method of study will be as follows: First, we will obtain expression for  $g_n(Z)$  which will be a function of  $\pi_2(n)$ , then we will use  $\pi_0$  and  $\pi_1$  to recalculate  $\pi_2(n)$  and obtain final expression for  $g_n(Z)$ .

The expression for  $g_n(Z)$  can be obtained directly as applying Theorem 3.1.2

$$g_n(Z) = \pi_2(n) \frac{1}{n^{m+1}} \left(\frac{\lambda}{2\mu}\right)^{m+1} (1 + o(1))$$

hence, as substituting the  $\pi_2(n) = \frac{1}{2} \frac{\lambda}{n} \frac{\lambda}{2\mu}$  in the relation above, we obtain

$$g_n(Z) = \frac{1}{n^{m+2}} \frac{1}{2} \left(\frac{\lambda}{2\mu}\right)^{m+2} (1 + o(1))$$

and from the expression for  $g_n(Z)$  we obtain, as setting  $\beta_n = n^{-m-2}$ , the parameter of exponential distribution as

$$A = \lambda \left(\frac{\lambda}{2\mu}\right)^{m+2}$$

Note that the result obtained for Theorem 5.1.1 is exactly the same as the result obtained above. We can conclude that the time of exit for both cases does not depend on the retrial rate if  $\nu_n \not\rightarrow 0$ .

## 5.2 $M/M/s/m$ system with retrials

Consider a Markov retrial queueing system of the type  $M/M/s/m$  with retrials. The system with  $s$  servers and  $m$  waiting places. The servers are independent and identical. Calls enter the system one at a time. On arrival, if one of the servers is free the customer will receive service immediately; otherwise, if there are free waiting positions the customer will join the queue waiting for service. On the other hand, if an arriving customer finds all servers and all of the

waiting positions occupied, the customer will be lost. Each waiting customer independently of others repeats its attempts for service after some random time. If at this time there is free server the customer is served, if all of the servers are busy again the call remains in the queue and repeats its attempts for service in the same way.

Let  $\lambda, \nu_n$  and  $\mu_n$  be given and  $\lambda$  be the input rate,  $\nu_n = n\nu$  be the rate for retrials for waiting customers and  $\mu_n = n\mu$  be the service intensity for each server. We will consider the cases:

**Case 1:**  $\nu_n = \nu$  (usual retrials) and  $\mu_n = n\mu$  (fast service) as  $n \rightarrow \infty$ .

**Case 2:**  $\nu_n = n\nu$  (fast retrials) and  $\mu_n = n\mu$  (fast service) as  $n \rightarrow \infty$ .

**Theorem 5.2.1** *For the system described above, under the assumption of fast service, independently of the initial state, the distribution of the normalized random variable  $n^{-s-m}\Omega_n(j, q)$  converges weakly to an exponentially distributed random variable*

$$\lim_{n \rightarrow \infty} \mathbf{P}\{n^{-s-m}\Omega_n(j, q) \geq t\} = \exp\{-At\}, \quad t > 0,$$

where

$$A = \lambda \left( \frac{\lambda}{s\mu} \right)^{m+s}$$

where  $\lambda$  is the input rate,  $\nu$  is the rate for retrial calls,  $n\mu$  is the service intensity.

**Proof:**

Let  $\Omega_n(j, q)$  be the time of first loss of a call given  $Q_n(0) = q$  and  $N_n(0) = j$  in a system where the rate of service is fast in the sense that  $\mu_n = n\mu$ . We study the asymptotic behavior of  $\Omega_n(j, q)$  as  $n \rightarrow \infty$ .

Consider a multicomponent process  $(N_n(t), Q_n(t))$  with state space  $S = \{0, 1, 2, \dots, s\} \times Z_+$ . The process  $(N_n(t), Q_n(t))$  forms a homogeneous MP in

continuous time and the state space for the process is in the form of

$$Z = \{(j, q), j = \overline{0, s}, q = \overline{0, m}\}$$

The rates of transitions for the process  $(N_n(t), Q_n(t))$  can be calculated and it can be seen that the subset  $Z$  forms monotone structure where at each fixed  $q = 0, 1, \dots, m$  the subset  $Z_q = \{(j, q), j = \overline{0, s}\}$  forms  $q$ -level.

In each state  $(j, q)$  the process spends an exponential time with parameter

$$\Lambda(j, q) = \begin{cases} \lambda + sn\mu & \text{if } j = s \\ \lambda + jn\mu + q\nu & \text{if } j \in (1, 2, \dots, s-1) \\ \lambda + q\nu & \text{if } j = 0 \end{cases}$$

Denote by  $\bar{\pi}_n(q) = (\pi_n(0, q), \pi_n(1, q), \dots, \pi_n(s, q))$  for  $q = s, s+1, \dots, s+m$ , the stationary distribution of the embedded Markov process for  $(N_n(t), Q_n(t))$  and let  $\pi_i, i = \overline{0, s}$  ( $\bar{\pi} = (\pi_0, \pi_1, \pi_2(n), \dots, \pi_s(n))$ ) be the stationary distribution for the states in  $Z_0$  to  $Z_s$ , where components  $\pi_0$  and  $\pi_1$  belong to the states in  $Z_0$  and  $\pi_k(n), k = \overline{2, s}$  belongs to the state in  $Z_i, i = \overline{1, s-1}$ , respectively. Note that  $Z_0$  in limit forms one essential class. The method of study will be as follows: First, we will obtain expression for  $g_n(Z)$  which will be a function of  $\pi_s(n)$ , then we will use  $\pi_0$  and  $\pi_1$  to recalculate  $\pi_s(n)$  and obtain final expression for  $g_n(Z)$ .

The expression for  $g_n(Z)$  can be obtained directly as applying matrix relation of Theorem 3.1.2

$$g_n(Z) = \pi_s(n) \frac{1}{n^{m+1}} \left(\frac{\lambda}{s\mu}\right)^{m+1} (1 + o(1))$$

hence, as substituting the

$$\pi_s(n) = \frac{1}{2} \frac{1}{n^{s-1}} \left(\frac{\lambda}{s\mu}\right)^{s-1}$$

in the above relation, we obtain

$$g_n(Z) = \frac{1}{n^{m+s}} \frac{1}{2} \left(\frac{\lambda}{s\mu}\right)^{m+s} (1 + o(1))$$

and from the expression for  $g_n(Z)$  we obtain, as setting  $\beta_n = n^{-m-s}$ , the parameter of exponential distribution as

$$A = \lambda \left( \frac{\lambda}{s\mu} \right)^{m+s}$$

**Theorem 5.2.2** *For the system described above, under the assumption of fast service and fast retrials, independently of the initial state, the distribution of the normalized random variable  $n^{-s-m}\Omega_n(j, q)$  converges weakly to an exponentially distributed random variable*

$$\lim_{n \rightarrow \infty} \mathbf{P}\{n^{-s-m}\Omega_n(j, q) \geq t\} = \exp\{-At\}, \quad t > 0,$$

where

$$A = \lambda \left( \frac{\lambda}{s\mu} \right)^{m+s}$$

where  $\lambda$  is the input rate,  $n\nu$  is the rate for retrial calls,  $n\mu$  is the service intensity.

### Proof

Let  $\Omega_n(j, q)$  be the time of first loss of a call given  $Q_n(0) = q$  and  $N_n(0) = j$  in a system where the rate of service and the rate of retrials are fast in the sense that  $\mu_n = n\mu$  and  $\nu_n = n\nu$ . The asymptotic behavior of  $\Omega_n(j, q)$  as  $n \rightarrow \infty$  is studied.

Consider a multicomponent process  $(N_n(t), Q_n(t))$  with state space  $S = \{0, 1, \dots, s\} \times Z_+$ . The process  $(N_n(t), Q_n(t))$  forms a homogeneous MP in continuous time and the state space for the process is in the form of

$$Z = \{(j, q), j = \overline{0, s}, q = \overline{0, m}\}$$

The rates of transitions for the process  $(N_n(t), Q_n(t))$  can be calculated and it can be seen that the subset  $Z$  forms monotone structure where at each fixed  $q = 0, 1, \dots, m$  the subset  $Z_q = \{(j, q), j = \overline{0, s}\}$  forms  $q$ -level.



In each state  $(j, q)$  the process spends an exponential time with parameter

$$\Lambda(j, q) = \begin{cases} \lambda + 2n\mu & \text{if } j = s \\ \lambda + jn\mu + qn\nu & \text{if } j \in (1, 2, \dots, s-1) \\ \lambda + qn\nu & \text{if } j = 0 \end{cases}$$

Denote by  $\bar{\pi}_n(q) = (\pi_n(0, q), \pi_n(1, q), \dots, \pi_n(s, q))$  for  $q = s, s+1, \dots, s+m$  the stationary distribution of the embedded Markov process for  $(N_n(t), Q_n(t))$  and let  $\pi_i, i = \overline{0, s}$  ( $\bar{\pi} = (\pi_0, \pi_1, \pi_2(n), \dots, \pi_s(n))$ ) be the stationary distribution for the states in  $Z_0$  to  $Z_{s-1}$  where components  $\pi_0$  and  $\pi_1$  belong to the states in  $Z_0$  and  $\pi_k(n), k = \overline{2, s}$  belongs to the state in  $Z_i, i = \overline{1, s-1}$ , respectively. Note that  $Z_0$  in limit forms one essential class. The method of study will be as follows: First, we will obtain expression for  $g_n(Z)$  which will be a function of  $\pi_s(n)$ , then we will use  $\pi_0$  and  $\pi_1$  to recalculate  $\pi_s(n)$  and obtain final expression for  $g_n(Z)$ .

The expression for  $g_n(Z)$  can be obtained directly as applying Theorem 3.1.2

$$g_n(Z) = \pi_s \frac{1}{n^{m+1}} \left( \frac{\lambda}{s\mu} \right)^{m+1} (1 + o(1))$$

hence, as substituting the

$$\pi_s = \frac{1}{2} \frac{1}{n^{s-1}} \left( \frac{\lambda}{s\mu} \right)^{s-1}$$

in the above relation, we obtain

$$g_n(Z) = \frac{1}{n^{m+s}} \frac{1}{2} \left( \frac{\lambda}{s\mu} \right)^{m+s} (1 + o(1))$$

and from the expression for  $g_n(Z)$  we obtain, as setting  $\beta_n = n^{-m-s}$ , the parameter of exponential distribution as

$$A = \lambda \left( \frac{\lambda}{s\mu} \right)^{m+s}$$

Note that the result obtained for Theorem 5.2.1 is exactly the same as the result obtained above. We can conclude that the time of exit for both cases does not depend on the retrial rate if  $\nu_n \not\rightarrow 0$ .

# Chapter 6

## SIMULATION RESULTS

Since the results obtained in the previous chapters are approximate results, we need to perform simulation analysis and see how the approximation technique works. The method of  $S$ -sets can be used in analysis of Markov systems, but it is much more harder to analyze non-Markov systems.

The simulation analysis is performed for the following two cases:

**Case 1:**  $M/M/1/m$  system with retrials.

System with single server and two waiting places ( $m = 2$ ) where customers arrive according to Poisson process with rate  $\lambda = 1$  customer per unit time, service rate is exponential with parameter  $\mu = 10$  customers per unit time and rate of retrial for each customer is exponential with parameter  $\nu = 2$  customers per unit time.

**Case 2:**  $M/G/1/m$  system with retrials.

System with single server and two waiting places ( $m = 2$ ) where customers arrive according to Poisson process with rate  $\lambda = 1$  customer per unit time, service rate is uniformly distributed between  $[0, 0.5]$  ( $U[0, 0.5]$ ) and rate of retrial for each customer is exponential with parameter  $\nu = 2$  customers per unit time.

## 6.1 Simulation of $M/M/1/m$ system with retrials

$M/M/1/m$  system with retrials is simulated to compare the approximation results with those obtained by simulation. The system can be described in the following manner. The system consists of a single server and two waiting positions. Customers arrive to the system according to Poisson process with rate  $\lambda = 1$  customer per unit time and if the server is free, an arriving customer starts service immediately and the service time will be exponential with parameter  $\mu = 10$  customers per unit time. If, upon arrival, the server is busy, the customer will join the retrial queue and will reapply for service after an exponential time with parameter  $\nu = 2$  customers per unit time. The capacity of the retrial queue is  $m = 2$ . If an incoming customer finds the server and all of the waiting positions full, the customer will leave the system forever.

We studied the behavior of the time of first lost customer in such a system and we know that this time under appropriate normalization and assumption of fast service weakly converges to an exponentially distributed random variable.

The results obtained from simulation analysis are shown in Table 6.1 and the comparison of simulated density and expected exponential density is shown in Figure 6.1

The result obtained using the formula of Theorem 4.1.1 for the case when  $m = 2$ ,  $\lambda = 1$ ,  $\mu = 10$ , and  $\nu = 2$  is  $A = \frac{3}{1600}$  (mean is 533.33). On the other hand the result obtained from simulation of the system of the type  $M/M/1/m$  with retrials, has mean of exponential distribution as 545.321 which is close to the value obtained by simulation.

The main difference between simulated value and approximation value is due to the small number of simulation trials (we performed the simulation only 160 times) and we chose  $n$  (scaling factor) to be of the order of 10 which is not very large.

3.184	3.856	6.632	8.542	11.784	14.479	15.547
16.33	18.925	19.389	19.698	22.308	23.568	25.888
32.639	36.458	40.058	42.551	44.034	53.194	57.836
61.592	64.324	73.765	74.918	87.834	88.785	91.555
93.325	97.602	105.196	116.552	119.658	122.638	127.25
132.059	141.831	142.072	142.448	167.38	175.316	185.186
189.639	191.017	210.931	215.414	219.685	222.958	238.644
241.469	241.578	242.372	243.1	246.614	246.637	251.363
257.863	258.813	267.711	268.779	272.965	274.054	286.813
291.216	296.262	297.397	301.743	317.148	317.501	318.593
324.294	324.623	329.766	348.053	385.591	386.477	389.934
391.959	421.904	422.092	424.142	428.915	436.713	436.841
457.197	459.011	469.757	469.853	478.066	486.408	507.896
512.016	512.131	526.21	527.106	530.806	531.176	536.914
537.842	553.689	565.988	567.425	569.083	587.823	588.622
618.293	628.375	655.6	664.216	666.202	666.299	676.752
688.114	698.099	709.676	723.063	728.492	741.744	754.546
783.471	797.205	843.845	844.282	854.884	874.513	879.755
892.09	897.025	905.43	906.496	978.087	996.985	1006.016
1009.595	1053.852	1064.096	1080.221	1090.025	1104.761	1106.961
1107.94	1115.806	1117.39	1200.67	1226.672	1237.699	1248.01
1274.386	1375.404	1488.44	1533.473	1613.755	1656.102	1737.921
1783.989	1798.627	1799.949	1802.216	2407.96	2553.084	

Table 6.1: Results of simulation of the time of first customer loss in the system of the type  $M/M/1/m$  with retrials where  $m = 2$ ,  $\lambda = 1$ ,  $\mu = 10$ , and  $\nu = 2$

If we increase the scaling factor  $n$  to be larger and would perform the simulation more than 10.000 times we would obtain much more better values by simulation (i.e, average of the simulated value would be closer to that obtained from approximate calculations).

As we noted previously, the simulation of rare events requires some special simulation techniques (see [41]). The simulation of rare events in Markov and non-Markov retrial queueing system can be a further research direction in this field. One can investigate for which values the scaling factor  $n$  can be accepted as large enough.

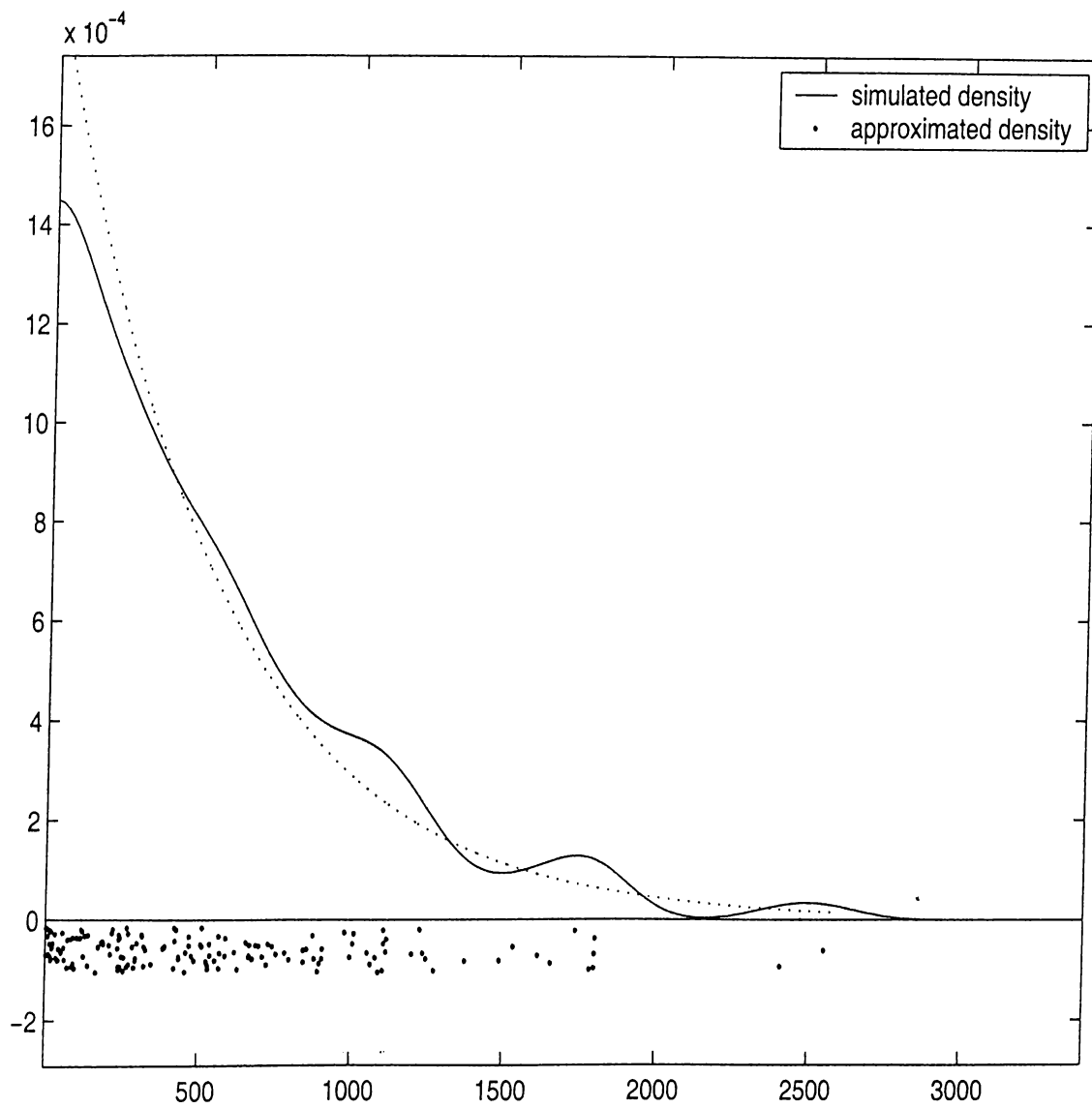


Figure 6.1: Approximated and simulated densities for the time of loss of first customer in a  $M/M/1/m$  system with retrials where  $m = 2$ ,  $\lambda = 1$ ,  $\mu = 10$   $\nu = 2$ .

## 6.2 Simulation of $M/G/1/m$ system with retrials

The method of  $S - sets$  can be used in the analysis of Markov systems but it is much more harder to study non-Markov systems using the method of  $S - sets$ , so we could not perform analysis for non-Markov systems. We can expect the time of exit from the system of the type  $M/G/1/m$  with retrials to be exponentially distributed random variable under the assumption of fast service. We performed simulation analysis for the non-Markov system  $M/G/1/m$  with retrials where the service time distribution is assumed to be uniformly distributed.

The system can be described in the following manner. The system consists of a single server and two waiting positions. Customers arrive to the system according to Poisson process with rate  $\lambda = 1$  customer per unit time and if the server is free, an arriving customer starts service immediately and the service time will be uniformly distributed in the interval  $[0, 0.5]$ . If, upon arrival, the server is busy, the customer will join the retrial queue and will reapply for service after an exponential time with parameter  $\nu = 2$  customers per unit time. The capacity of the retrial queue is  $m = 2$ . If an incoming customer finds the server and all of the waiting positions full, the customer will leave the system forever.

The results obtained from simulation for the cases when service rates are uniformly distributed on the intervals  $U[0, 0.5]$  and  $U[0, 0.2]$  are shown in Table 6.2 and Table 6.3, respectively. Also, comparison of simulated density and exponential density are shown in Figure 6.2 and Figure 6.3 for both cases. Simulated parameter for  $U[0, 0.2]$  is  $A = 1/780.42$  and simulated parameter for  $U[0, 0.5]$  is  $A = 1/91.55$ .

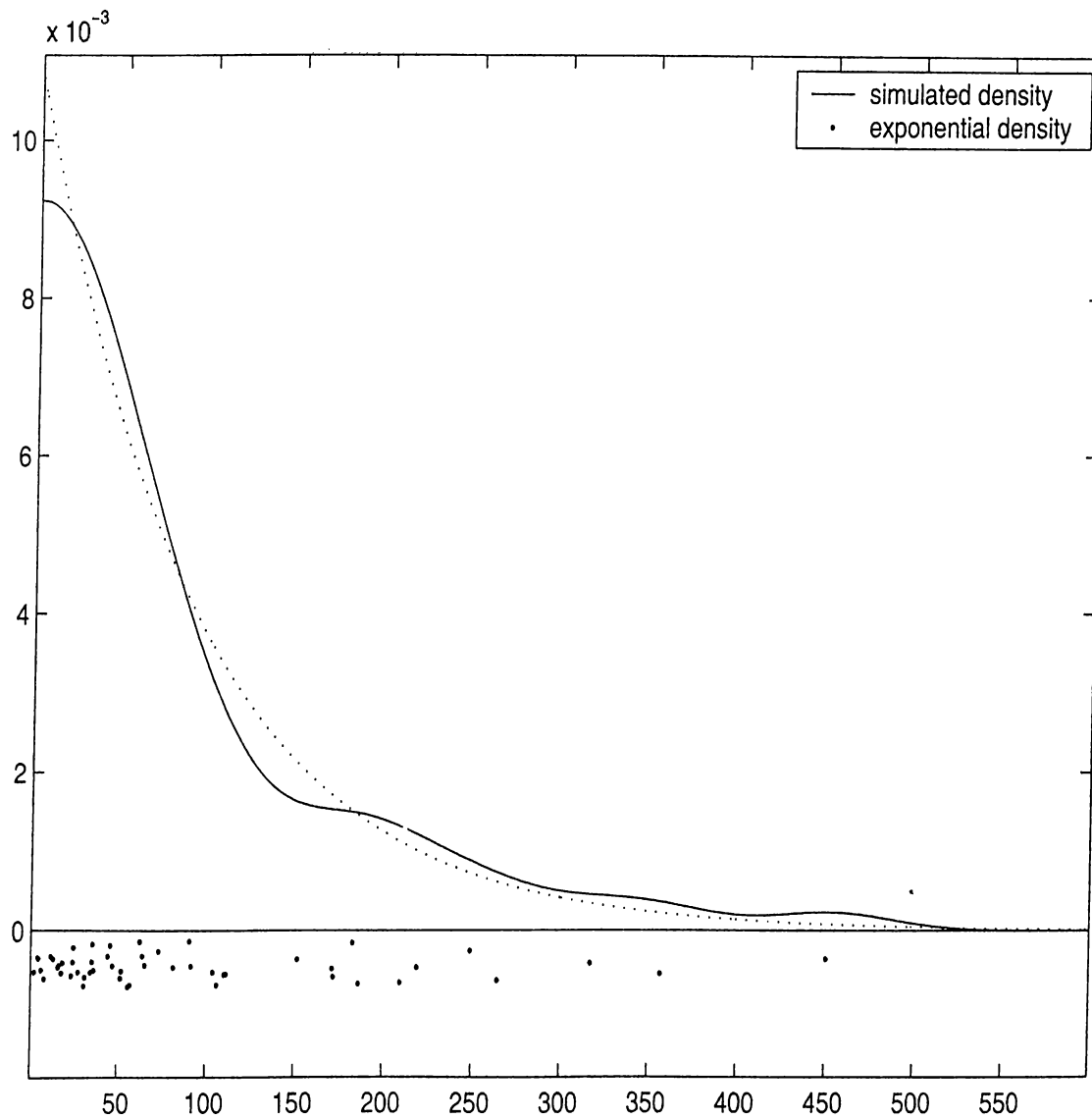


Figure 6.2: Approximated and simulated densities for the time of loss of first customer in a  $M/G/1/m$  with retrials where  $m = 2$ ,  $\lambda = 1$ ,  $\nu = 2$  and service times are uniformly distributed on the interval  $[0, 0.5]$ .

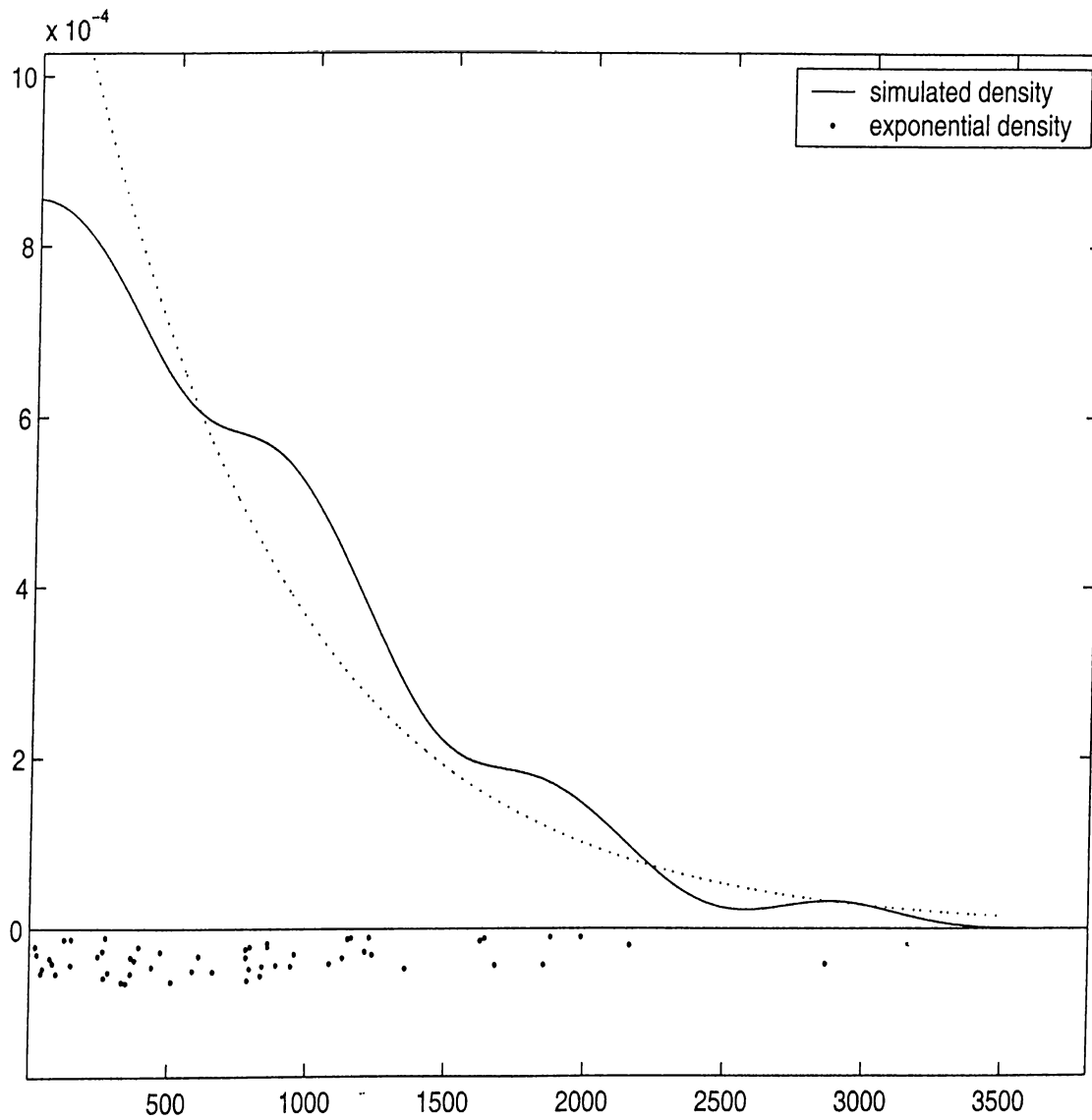


Figure 6.3: Approximated and simulated densities for the time of loss of first customer in a  $M/G/1/m$  with retrials where  $m = 2$ ,  $\lambda = 1$ ,  $\nu = 2$  and service times are uniformly distributed on the interval  $[0, 0.2]$ .



2.566	4.483	5.029	6.446	8.221
12.045	13.367	16.1	16.452	18.009
18.645	23.668	24.589	27.418	30.868
31.286	34.542	35.383	35.791	36.382
44.363	45.674	47.026	51.576	51.977
55.824	56.949	62.428	63.923	65.308
73.046	81.398	90.43	91.5	103.877
106.182	110.241	111.352	151.586	171.074
171.849	182.604	186.054	209.688	219.172
248.999	264.33	317.296	356.875	450.409

Table 6.2: Results of simulation of the time of first customer loss in the system of the type  $M/G/1/m$  with retrials where  $m = 2$ ,  $\lambda = 1$ ,  $\nu = 2$  and service times are uniformly distributed on the interval  $[0, 0.5]$ .

22.291	28.267	42.901	48.42	74.725
83.572	96.85	125.368	148.956	150.6
245.858	262.708	266.762	272.47	282.066
330.856	346.204	363.184	363.95	377.421
392.372	439.792	471.117	510.087	586.226
608.23	659.679	777.863	778.604	783.319
791.318	792.812	831.125	836.562	855.93
856.505	887.199	941.248	954.121	1081.796
1129.714	1148.186	1159.879	1209.135	1224.8
1235.548	1354.681	1625.982	1641.299	1679.173
1853.042	1877.712	1987.76	2161.174	2865.675

Table 6.3: Results of simulation of the time of first customer loss in the system of the type  $M/G/1/m$  with retrials where  $m = 2$ ,  $\lambda = 1$ ,  $\nu = 2$  and service times are uniformly distributed on the interval  $[0, 0.2]$ .

# Chapter 7

## CONCLUSION

This thesis investigates the asymptotic behavior of the time of first loss of customer in retrial queueing models of the single server and multiple server types. We analyze the systems under two different assumptions: a model where service is considered to be fast and a model where both service and retrials are considered to be fast.

We used the method of  $S$  – sets to prove that the time of first customer loss from the given Markov system, under appropriate scaling, weakly converges in distribution to an exponentially distributed random variable.

We analyzed single server retrial queueing systems of various types

- Single-server retrial queueing system where service is considered asymptotically fast and both the rate of incoming customers and the rate of retrials are of usual orders. Also the results for the case when rates depend on the size of the queue were considered.
- Single-server retrial queueing system where both service and retrials are considered asymptotically fast and the rate of incoming customers is of usual order. Also the results for the case when rates depend on the size of the queue were considered.

- Single-server retrial queueing system operating in Markov environment where the service is considered asymptotically fast. Also the rates in the model depend on the size of the queue and on the state of the additional Markov environment.

An exponential approximation for the time of loss of first customer was proved and the parameter of exponential distribution was derived for all of the cases described above.

We also considered the multiple-server retrial queueing systems of various types

- Two-server retrial queueing system where service is considered asymptotically fast and the rate of incoming customers and the rate of retrials are of usual orders. We also considered the case where both service rate and rate of retrials are considered asymptotically fast and rate of incoming customers is of usual order.
- $s$ -server retrial queueing system where service is considered asymptotically fast and both the rate of incoming customers and the rate of retrials are of usual orders. We also considered the case where both service rate and the rate of retrials are considered asymptotically fast and the rate of incoming customers is of usual order.

An exponential approximation for the time of loss of first customer was proved and the parameter of exponential distribution was derived for all of the cases described above.

Table 7.1 and Table 7.2 summarizes the results derived in Chapter 4 and 5 for the time of loss of first customer in retrial queueing system with waiting positions under some different assumptions.

### Future Research Directions

- Single server retrial queueing model where server is subject to breakdowns (unreliable server) with any combination of fast service and fast repairs.
- Also multiple server models with servers subject to breakdowns can be considered.
- Non-Markov retrial queueing systems with fast service. We can consider the case where arrivals are non-Markov and service is exponential as well as the case where service is non-Markov and arrivals are exponential and asymptotic behavior for the time of loss of first customer can be considered.
- Another direction can be the simulation of the retrial queueing systems. Systems for which asymptotic analysis is not possible can be simulated and various characteristics can be obtained.

System (assumptions)	Distribution of time of loss of first customer	Parameter	Order
<i>M/M/1/m with retrials (fast service)</i>	Exponential	$A = \frac{\lambda \rho^{m+1}}{m! v^m} \prod_{k=1}^m (\lambda + kv)$	$n^{-m-1}$
<i>M/M/1/m with retrials (fast service &amp; fast retrials)</i>	Exponential	$A = \lambda \left( \frac{\lambda}{\mu} \right)^{m+1}$	$n^{-m-1}$
<i>M/M/1/m with retrials (fast service and dependence on queue length)</i>	Exponential	$A = \lambda(0) \frac{1}{m!} \prod_{k=0}^m \frac{\lambda(k)}{\mu(k)} \prod_{k=0}^{m-1} \frac{\lambda(k) + (k+1)v}{v(k+1)}$	$n^{-m-1}$
<i>M/M/1/m with retrials (fast service &amp; fast retrials and dependence on queue length)</i>	Exponential	$A = \lambda(0) \prod_{k=0}^m \frac{\lambda(k)}{\mu(k)}$	$n^{-m-1}$
<i>M/M/1/m with retrials (operating in Markov environment)</i>	Exponential	$A = \bar{\pi} \Lambda(0) G(0) (I - B(1) - \Lambda(1))^{-1} (I - B(1)) G(1) \dots \dots G(m-1) (I - B(m) - \Lambda(m))^{-1} (I - B(m)) G(m) \bar{1}$	$n^{-m-1}$

Figure 7.1: Summary of the results for single server retrial queueing models

System (assumptions)	Distribution of time of loss of first customer	Parameter	Order
<i>M/M/2/m with retrials (fast service)</i>	Exponential	$A = \lambda \left( \frac{\lambda}{2 \mu} \right)^{m+2}$	$n^{-m-2}$
<i>M/M/2/m with retrials (fast service &amp; fast retrials)</i>	Exponential	$A = \lambda \left( \frac{\lambda}{2 \mu} \right)^{m+2}$	$n^{-m-2}$
<i>M/M/3/m with retrials (fast service)</i>	Exponential	$A = \lambda \left( \frac{\lambda}{3 \mu} \right)^{m+3}$	$n^{-m-3}$
<i>M/M/3/m with retrials (fast service &amp; fast retrials)</i>	Exponential	$A = \lambda \left( \frac{\lambda}{3 \mu} \right)^{m+3}$	$n^{-m-3}$
<i>M/M/s/m with retrials (fast service)</i>	Exponential	$A = \lambda \left( \frac{\lambda}{s \mu} \right)^{m+s}$	$n^{-m-s}$
<i>M/M/s/m with retrials (fast service &amp; fast retrials)</i>	Exponential	$A = \lambda \left( \frac{\lambda}{s \mu} \right)^{m+s}$	$n^{-m-s}$

Figure 7.2: Summary of the results for multiple server retrial queueing models

# Bibliography

- [1] A. M. Aleksandrov, A queueing system with repeated orders, *Engineering Cybernetics Rev.* **12**, 3(1974)1.
- [2] V. V. Anisimov, Limit distributions of functionals of a semi-Markov process given on a fixed set of states up to the time of first exit, *Soviet Math. Dokl.* **11** (1970), No.4, 1002-1004.
- [3] V. V. Anisimov. Asymptotic consolidation of the states of random processes, *Cybernetics* **9** No.3 (1973) 494-504.
- [4] V. V. Anisimov, Limit theorems for sums of random variables in an array of sequences defined on a subset of states of a Markov chain up to the exit time, *Theor. Probability and Math. Stat.* **4** (1974) 1.
- [5] V. V. Anisimov. Limit theorems for random processes and their applications to discrete summation schemes, *Teoria Veroyatnostey i Primenen.* **20** No.3, (1975) 692-694. English translation in *Theor. Probab. Appl.* **20**, (1975).
- [6] V. V. Anisimov. Switching processes. *Cybernetics* **13** No.4 (1977) 590-595.
- [7] V. V. Anisimov. Limit theorems for switching processes and their applications, *Cybernetics* **14** No.6 (1978) 917-929.
- [8] V. V. Anisimov, Inequalities in Markov approximation of lumped processes, *Probability Theory and Mathematical Statistics*. Proc. 4-th Vilnius Conf., USSR, 1985, VNU Science Press, The Netherlands. Vol. **1**, (1988).

- [9] V. V. Anisimov, O. K. Zakusilo, and V. S. Dontchenko, The elements of queueing theory and asymptotic analysis of systems, *Publ. "Visca Scola" Kiev* p.248 (1987) (in Russian).
- [10] V. V. Anisimov and S. G. Pushkin, Limit theorems and proximity estimates for summation schemes on Markov chains, *Theory of Probability and Mathematical Statistics*, **37** (1988).
- [11] V. V. Anisimov, Random processes with discrete component. Limit theorems, *Publ. Kiev Univ.* p.184 (1988a) (in Russian).
- [12] V.V. Anisimov and J. Sztrik. Asymptotic analysis of some controlled finite-source queueing systems, *Acta Cybernet.* **9** (1989), No.1, 27-38.
- [13] V. V. Anisimov and J. Sztrik. Asymptotic analysis of some complex renewable system operating in random environment, *European Journal of Operations Research* **41**(1989b), 162-168.
- [14] V. V. Anisimov and J. Sztrik. Reliability analysis of a complex renewable system with fast repair, *J. of Information Processing and Cybernetics*, EIK, Berlin, **25**, No. 11/12, (1989c) 573.
- [15] V. V. Anisimov. Switching processes: Averaging Principle, Diffusion Approximation and Applications, *Acta Applicandae Mathematicae*, Kluwer **40** (1995) 95-141.
- [16] V. V. Anisimov, Asymptotic analysis of switching queueing systems in conditions of low and heavy loading, *Matrix-Analytic Methods in Stochastic Models*, Eds. S.R. Chakravorthy and A.S. Alfa, Lecture notes in Pure and Applied Mathematics Series, Marcel Dekker, Inc., **183** (1996), 241-260.
- [17] V. V. Anisimov. Asymptotic merging of states in hierarchical stochastic models and applications in queueing networks, *Advances in Computer and Information Sciences '98*, Editors, U. Gudukbay et al., IOS Press (1998).



- [18] V.V. Anisimov. Averaging methods for transient regimes in overloading retrieval queueing systems, *Mathematical and Computer Modelling* **30**(1999) 65-78.
- [19] V. V. Anisimov. Averaging methods for switching queueing networks in asymptotically consolidated environment, *Proc. of 11th European Simulation Conference ESS'99*, Erlangen, Germany, Oct. 26-28, (1999), 682-686.
- [20] V. V. Anisimov. Asymptotic analysis of reliability for switching systems in light and heavy traffic conditions, *Recent Advances in Reliability Theory: Methodology, Practice and Inference*, Eds: N. Limnios, M. Nikulin, Birkhauser Boston Inc., (2000) (forthcoming)
- [21] A. Bobbio and K. S. Trivedi, An aggregation technique for the transient analysis of stiff Markov chains, *IEEE Transactions on Computers*, C-35 **9**, (1986) 803-814.
- [22] A. A. Borovkov, Asymptotic Methods in Queueing Theory, *John Wiley and Sons Ltd.*, 1984.
- [23] Q. H. Choo and B. Conolly, New results in the theory of repeated orders queueing systems, *J. Appl. Prob.* **16** (1979) 631.
- [24] J. W. Cohen, Basic problems of telephone traffic theory and the influence of repeated calls, *Philips Telecom. Rev.* **18**, 2(1957) 49.
- [25] G. I. Falin, Aggregate arrival of customers in one-line system with repeated calls, *Ukrainian Math. J.* **28** (1976) 437 (in Russian).
- [26] G. I. Falin, A single-line system with secondary orders, *Engineering Cybernetics Rev.* **17** 2(1979) 76 (in Russian).
- [27] G. I. Falin, Not completely accessible schemes with allowance for repeated calls, *Engineering Cybernetics Rev.* **18**, 5(1980) 56 (in Russian).
- [28] G. I. Falin, Investigation of weakly loaded switching systems with repeated calls, *Engineering Cybernetics Rev.* **19**, 3(1981) 69.

- [29] G.I. Falin. Asymptotic investigation of completely accessible switching systems with a high rate of repetition of blocked calls, *Vestnik Moskov.Univ.Ser.I Math.Mekh.* **6** (1984) 49-53, 111 (in Russian).
- [30] G.I. Falin. On the waiting-time process in a single-line queue with repeated calls, *Journal of Applied Probability* **23** (1986) 185-192.
- [31] G.I. Falin. Multichannel queueing systems with repeated calls under high intensity of repetition, *Journal of Inform. Processes. Cybernet.* **23** (1987) 37-47 (in Russian).
- [32] G. I. Falin, On a multiclass batch arrival retrial queue, *Adv. Appl. Prob.* **20** (1988) 483-487.
- [33] G. I. Falin and J. G. C. Templeton, *Retrial Queues*, ChapmanHall (1997)
- [34] G. Falin, A survey of retrial queues, *Queueing systems* **7**(1990) 127-168.
- [35] O. Hashida and K. Kawashima, Buffer behavior with repeated calls, *Electronics and Communication in Japan* **62-B**, 3(1979) 27.
- [36] G. L. Jonin and Y. Y. Sedol, Investigation of telephone systems in the case of repeated calls, *Latvian Mathematical Yearbook* **7** (1970) 71.
- [37] J. Keilson, J. Cozzolino and H. Young, A service system with unfilled requests repeated, *Oper. Res.* **16** (1968) 1126.
- [38] V.S. Korolyuk and A.F. Turbin, *Mathematical Foundations of Phase Consolidations of Complex Systems*, Publ. "Naukova Dumka", Kiev (1978) (in Russian).
- [39] L. Kosten, On the influence of repeated calls in the theories of probabilities of blocking, *De Ingenieur* **59** (1947) 1.
- [40] I.N. Kovalenko, Rare Events Analysis in the Estimation of Systems Efficiency and Reliability, Publ. "Sov. Radio", Moscow (1980) (in Russian).
- [41] I. N. Kovalenko, Rare events in queueing systems, A survey, *Queueing Systems* **16** (1994) 1-49.

- [42] V. G. Kulkarni, Expected waiting times in a multiclass batch arrival retrial queue, *J. Appl. Prob.* **23** (1986) 144.
- [43] P. Le Gall, The repeated call model and the queue with impatience, *Proc. Third Int. Seminar on Teletraffic Theory*, Moscow (1984) 278-289.
- [44] L. Lipsky, Queueing Theory, A Linear Algebraic Approach, *Macmillian Publishing Company*, USA, 1992.
- [45] M. F. Neuts Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach. *Dover Publ.* 1995. (First Published in 1981 by Johns Hopkins University Press).
- [46] N. U. Prabhu, Foundations of Queueing Theory, *Kluwer Academic Publishers* ( International Series in Operations Research and Management Science,7), 1997.
- [47] G. E. Ridout, A study of retrial queueing systems with buffers, M.A.Sc. Thesis, Department of Industrial Engineering, University of Toronto (1984).
- [48] J. Riordan, Stochastic Service Systems, (Wiley, New York, 1962).
- [49] A.D. Soloviev, Asymptotic behavior of the first occurrence time of a rare event in a regenerative process, *Izv. Akad. Nauk. SSSR Tekhn. Kibern.*, **6** (1971) 79 (in Russian).
- [50] S. N. Stepanov, Integral equilibrium relations of non-full-access systems with repeated calls and their applications, *Prob. Inf. Trans.* **16**, 4(1980) 88 (in Russian).
- [51] S. N. Stepanov, Probabilistic characteristics of an incompletely accessible multi-phase service system with several types of repeated calls, *Problems of Control and Information Theory* **10**, (1981) 387 (in Russian).
- [52] S. N. Stepanov, Asymptotic formulae and estimations for probabilistic characteristics of full-available group with absolutely persistent subscribers, *Problems of Control and Information Theory* **12**, 5(1983) 361 (in Russian).

- [53] S. N. Stepanov, Probabilistic characteristics of an incompletely accessible service system with repeated calls for arbitrary values of subscriber persistent function, *Problems of Control and Information Theory* **13**, 2(1984) 69 (in Russian).
- [54] S. N. Stepanov and I. I. Tsitovich, The model of a full-available group with repeated calls and waiting positions in the case of extreme load, *Problems of Control and Information Theory* **14**, 1(1985) 25 (in Russian).
- [55] J. Sztrik and D. Kouvatsos. Asymptotic analysis of a heterogeneous multiprocessor system in a randomly changing environment, *IEEE Transactions on Software Engineering* **17** (1991), No.10, 1069-1075.
- [56] J. Sztrik. Asymptotic analysis of a heterogeneous renewable complex system with random environments, *Microelectronics and Reliability* **32** (1992) 975-986.
- [57] R. I. Wilkinson, Theories for toll traffic engineering in the U.S.A., *Bell System Tech. J.* **35**, (1956) 421.
- [58] T. Yang and J. G. C. Templeton, A survey on retrial queues *Queueing Systems* **2** (1987) 203-233.

## VITA

Mümin Kurtuluş was born on March 5, 1976 in Kırçali, Bulgaria. He completed his primary education in Kırçali, Bulgaria. In 1989, his family moved to Istanbul, Turkey and he received his high school diploma from Yeşilköy 50. Yıl Lisesi, Istanbul, Turkey. He has received his Bachelor of Science Degree from the Department of Physics, Koc University, Istanbul, Turkey. In September 1998, he joined the Department of Industrial Engineering at Bilkent University, Ankara, Turkey as a research assistant. From that time to the present, he worked with Professor Vladimir V. Anisimov for his master's thesis at the same department.