Order Number 9009597

Sample-average analyses of some generalizations of the M/G/1 queue

Li, Jingwen, Ph.D.

The University of Texas at Dallas, 1989

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

Sample–Average Analyses of Some Generalizations of the M/G/1 Queue

by

Jingwen Li, B.S., M.S.

DISSERTATION

Presented to the Faculty of

The University of Texas at Dallas

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY IN MANAGEMENT SCIENCE

THE UNIVERSITY OF TEXAS AT DALLAS

November 1989

Sample–Average Analyses of Some Generalizations of the M/G/1 Queue

APPROVED BY SUPERVISORY COMMITTEE:

Shun-Chen Niu

Shun–Chen Niu (Chairman)

Ramaswamy Chandrasekaran

George Kimeldorf

Ping Yang

ACKNOWLEDGEMENTS

I would like to thank Professor R. Chandrasekaran, Professor S.C. Niu, Professor G. Kimeldorf, and Professor P. Yang for serving in my dissertation committee. I am especially indebted to Professor S.C. Niu, my academic advisor, for his guidance and constant support throughout the course of my thesis work. Special thanks are due to Ms. Janice Jantz and Ms. Abbie Bailey for their constant help during the preparation of this dissertation.

I am deeply indebted to my wife Jingyu Wang for her devotedness and constant encouragement during the past years.

Sample–Average Analyses of Some Generalizations of the M/G/1 Queue

Publication No._____

Jingwen Li, Ph.D.

The University of Texas at Dallas, 1989

Supervising Professor: Shun–Chen Niu

This dissertation employs sample–average methods to study two important generalizations of the M/G/1 queue: M/G/1 queues with modified services at the beginning of busy periods; and finite–capacity M/G/1/K queues with more general arrival processes. All of our results are given in explicit, transform–free form.

In the first generalization, we consider a variety of M/G/1 models that have modified services at the beginning of busy periods, such as M/G/1 queues with exceptional first service, M/G/1 queues with set–up times, M/G/1 queues with server vacations, and M/G/1 queues under D–policy or N–policy. We study, via the preemptive–resume–last–in–first–out queue discipline, sample–average behavior of cumulative work in these systems, as observed by arriving customers. We derive waiting–time distributions of customers in these models. We also establish decomposition results for the GI/G/1 queue with server vacation. We further show that the decomposition result holds for more general vacation models. All results for these models are in fact given in the context of

GI/G/1 queues, except for M/G/1 queues under D–policy or N–policy. Further more, our analyses are valid even when the system has combined feature of these modified services at the beginning of busy periods. As an application of some of our results, we compare control policies under linear cost assumptions.

In the second generalization, the arrival process we consider has the Markovian property and is governed by an "underlying" continuous–time Markov chain. Such arrival processes generalize several well–known point processes, such as the Markov modulated arrival process, the continuous–time Markov chain generated arrival process, any "phase–type" arrival process, and superpositions of these processes. For this model, we derive formulas for the long–run average joint behavior of queue length and remaining service time of the customer (if any) in service, over customer arrival epochs. We also discuss in detail computational issues in connection with these formulas. In particular, we describe very efficient computational procedures when the service–time distribution is either generalized hyperexponential or Erlang.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Queueing theory was developed primarily to provide models to predict behavior of systems that provide service for randomly arising demands. It originated as a very practical subject. The pioneering investigator was the Danish mathematician A.K. Erlang who, in 1909, published a book titled *The Theory of Probabilities and Telephone Conversations*, in which he observed that a telephone system was generally characterized by Poisson input, exponential or constant holding (service) time, and multiple channels (servers). Based on this observation, he built a mathematical model to determine the optimal number of telephone lines to handle prescribed incoming call frequencies. This mathematical model is known today as a birth–and–death process, and it continues to serve as the backbone mathematical model for the study of telephone systems.

Telecommunication engineering remains a principal area of research and application of queueing theory today. Equipped with high–speed digital computers, modern telecommunication networks become increasingly more complicated. At the same time, numerous other applications of queueing theory have also been discovered in production line planning, machine repair scheduling, toll booths and taxi stands handling, inventory controlling, air traffic controlling, and computer system designing. Along with the increase in the areas of applications of queueing theory, the complexity of the mathematical models developed for various problems also increases rapidly. The birth–and–death process used by Erlang is therefore no longer sufficient to meet the increasing needs of modeling more general physical queueing systems, a situation that naturally leads to the introduction of more sophisticated and also more complicated mathematical models in queueing theory. It also brings about, inevitably, many new mathematical complications which queueing theorists must address.

As a direct consequence of extensive use of advanced mathematics, research results in modern queueing literature have become increasingly more difficult to interpret. For example, many research results in queueing literature are given either in Laplace–Stieltjes–transform form, or in terms of roots determined by some analytic equations on the complex plane. The behaviors of these roots are not well understood. To mathematicians, results in such form are considered acceptable, because the solutions of the problems concerned are uniquely determined. But to those who need to apply queueing theory to managerial decision making, it is extremely important that results be presented in implementable form. This concern makes results given in Laplace–Stieltjes transform or in terms of roots determined by analytic equations on the complex plane no longer acceptable, for at least the following two reasons: (1) such results do not provide clear physical connections between the underlying structures of the models under consideration and the obtained final results, because of the so called "Laplacian curtain", or lack of understanding of the behavior of those roots; (2) although results in such form are uniquely determined, they do not suggest procedures for numerical computation. For instance, Laplace–Stieltjes–transform formulas for waiting–time distributions in queueing models do not in general provide schemes for computing the explicit distributions, because inversion of Laplace–Stieltjes transforms requires specially designed analytic integration on the complex plane. In extreme cases, the inversions necessary for obtaining explicit formulas could be as difficult as solving the original problem. In summary, modern queueing theory should mean not only building realistic mathematical models, but also aiming at meaningful and implementable final results. Based on this reasoning, we prefer to employ solution methods that lead to results in forms that not only reflect explicitly physical attributes of the models, but also are ready for numerical computation. Sample–average methods are ideal with respect to these objectives. Such methods work directly on typical realizations of the processes under study. They provide not only direct physical interpretations for the

obtained results along every step of the analyses, but also natural procedures for numerical computation. This is the primary reason for our use of sample–average methods in this dissertation.

This dissertation is application oriented. Our purpose is to analyze delay distributions of customers in several generalized M/G/1 queueing systems that arise naturally in applications. The notation M/G/1 here stands for Markovian (or Poisson) arrival process, general (usually renewal) service process, and one server. The generalized M/G/1 queueing models we analyze can be classified into two categories: (1) M/G/1 queues with modified services at the beginning of busy periods, and (2) M/G/1 finite–capacity queues with more general arrival processes.

There are five different models in the first category. They are: (1) M/G/1 queues with exceptional first services in busy periods; (2) M/G/1 queues with set–up times; (3) M/G/1 queues with server vacations; (4) M/G/1 queues under certain control policies; (5) models of type (1), (2), and (4) with server vacations. All of these models arise frequently in applications.

Among these models, the M/G/1 queue with exceptional first service in each busy period has been studied previously by, for example, Welch [1964] and Itzhak, Maxwell and Miller [1965]. All the previous results for this model are given in Laplace–Stieltjes–transform form. In this dissertation, we will derive explicit, sample–average results for both randomly selected customers as well as those who do not initiate busy period in this model. Our methods of analyses are, in fact, valid in more general GI/G/1 settings. These explicit results serve as basic tools for analyzing other models in this category.

In M/G/1 queues with set–up times, the server spends a random set–up time before starting service at the beginning of each busy period. Because of this delay, the waiting time of customers in the system increases stochastically. This model was first proposed by

Yadin and Naor [1963]. Doshi [1985] studied the waiting–time distribution of customers in this model in the GI/G/1 setting (where GI stands for "general renewal arrival process"). A particularly interesting result of Doshi is that the waiting time of customers in this system is a sum of two random variables, one of which is the waiting time of customers in a standard GI/G/1 queue, and the other follows a distribution that is given in a complicated transform form. In the next chapter, we will give a sample–average proof of his result, as well an explicit expression for the waiting–time distribution of customers in this system.

The vacation model operates slightly differently. Specifically, whenever the server finishes all the work in the system, he takes a vacation. If the server returns from a vacation and finds at least one customer in the system, he starts service immediately; otherwise, he takes another vacation. This model was early studied by Keilson [1962], Gaver [1962], Skinner [1967], Cooper and Murray [1969]; and then studied subsequently by many other queueing analysts. They showed that the waiting time of customers in the system is decomposed into two random variables, one of which is the same as the waiting time of customers in a standard M/G/1 queue, and the other is distributed as the equilibrium excess (or forward recurrence time) of a typical vacation time. Doshi [1985] generalized this earlier decomposition results for the M/G/1 vacation model to the context of GI/G/1. In this dissertation, we will give an intuitive proof of Doshi's result and extend it to more general settings.

M/G/1 queues under different control policies can be viewed as a class of M/G/1 queues with modified services at the beginning of busy periods. They originated in problems of optimal system control in queueing theory. The primary objective of system control is to determine, for a given policy, the optimal values of control variables (to be specified later) so that the long–run average cost of the system is minimized. One of the basic control policies is the N–policy, which was proposed and studied by Yadin and Naor [1963] and further analyzed by Heyman [1968]. Under this policy, the system is turned on

whenever the number of customers in the system reaches a predetermined integer N and is turned off when the server completes servicing all waiting customers. Heyman [1977] also proposed and studied a related T–policy, which differs from the N–policy in that instead of waiting for the number of customers in the system to reach N, the system is turned on after a constant time interval, of length T, since it was last turned off; if there is no customer in the system when it is turned on, the system is turned off immediately. Balachandran [1973] and Balachandran and Tijms [1975] studied another related policy, called the D–policy, which is similar to the N–policy but it turns the system on whenever the cumulative work in the system exceeds a predetermined threshold of size D. Balachandran and Tijms [1975] conjectured that if the waiting cost charged to the system is a linear function of the time–average workload, then the D–policy is better than the N–policy. This conjecture was later proved by Boxma [1976]. For more detailed description and discussions of these policies, please refer to the papers cited above, as well as to Crabill, Gross and Magazine [1977] and to Cooper [1981], pp. 343–253.

An important observation concerning queueing research in the area of system control is that systematic studies of the probabilistic behavior of M/G/1 queues under various control policies do not exist. This is primarily due to the difficulties involved in solving these systems by traditional methods, especially in the case of the M/G/1 queue under the D–policy. This situation forces queueing analysts to impose restrictive assumptions on the cost functions that are used in their analysis of optimal control so that only minimal information about the systems is required. Clearly, restrictive assumptions apply only to limited cases. General analysis of system control does require detailed information on the probabilistic behavior of these queueing systems (waiting–time distributions, in particular). In this dissertation, we will derive waiting–time distributions of customers in M/G/1 queues under the D–policy and the N–policy, using sample–average methods.

The basic idea behind our analyses of all these M/G/1 queues with modified services at the beginning of busy periods is to study the behavior of workloads in the systems as seen by arriving customers via the preemptive–resume–LIFO (last–in, first–out) queue discipline. In this queue discipline, the service requested by the first customer in a busy period will always be finished the last. If a customer arrives at times epoch t, then the total workload in the system (if any) contributed by those who entered the system after the start of the busy period in progress is not affected in any way by the customer who initiated the busy period. If we interpret the set–up times or the remaining vacation times as services brought in by "artificial" customers who initiate busy periods, then, this observation, together with the fact that the preemptive–resume LIFO queue discipline is work–conserving, leads to direct explanation of the decomposition results for M/G/1 queues with set–up times or with server vacations.

We now discuss M/G/1 systems in our second category of generalizations, namely M/G/1 finite–capacity queues with more general arrival processes. An important aspect of these models is the constraint on system capacity. This constraint usually causes considerable analytic difficulties when one attempts to apply traditional methods for infinite–capacity queues to the analyses of finite–capacity systems. Consider the M/G/1 queue, for example. For this model, it is typical to start the analysis by analyzing a Markov chain embedded at customer–departure epochs. If the capacity is not bounded above, the probabilistic behavior of cumulative work brought in the system by future arrivals will not depend on the current state of the embedded Markov chain (i.e. the queue length). This property, which is absent in finite–capacity queues, offers considerable analytic convenience to queueing analysts when they study infinite–capacity queues. In reality, almost all queues have limited capacities. Only for queues whose capacities are so large that the probability that an arriving customer sees the system full is negligible can the infinite–capacity approximation be justified. In this dissertation, we will develop

unified methods for the analyses of both infinite– and finite–capacity queues.

Another important aspect of models in this category is that we allow the arrival process to the system to be much more general than Poisson. There are two basic motivations for making such generalizations of the M/G/1/K queue: (1) we want to allow for dependent interarrival times so that the resulting system is more realistic; (2) we also want the arrival process to be versatile enough so that it could accommodate a large variety of applications. The particular class of arrival processes we consider will be "Markovian" arrival processes that are "controlled" or "governed" by an underlying continuous–time Markov chain. We will call this type of arrival processes C–processes. There are many point processes fitting this description, such as: (1) continuous–time–Markov–chain generated arrival processes, (2) doubly stochastic Poisson processes (also called "Markov modulated processes"; see, for example, Rogterschot and deSmit [1986]), (3) "phase–type" renewal processes, and (4) superpositions of these arrival processes. In Chapter 3 we will give additional explanations for considering such arrival processes and we will analyze the long–run average joint behavior of queue length and remaining service time of the customer, if any, in service, over customer arrival epochs, in the C/G/1/K queue. Our results provide explicit state information needed for the derivation of many other quantities of interest in this model, notably the waiting–time distribution. All of our results are given in explicit, transform–free form. We also discuss in detail computational issues related to these results. In particular, we show that if the service–time distribution is either generalized hyperexponential or Erlang, then the computational complexity of our schemes can be reduced significantly.

CHAPTER 2

ANALYSES OF M/G/1 QUEUES WITH MODIFIED SERVICES AT THE BEGINNING OF BUSY PERIODS

2.1 INTRODUCTION

The models we study in this chapter are: (1) M/G/1 queues with exceptional first services in busy periods; (2) M/G/1 queues with set–up times; (3) M/G/1 queues with server vacations; (4) M/G/1 queues under certain control policies; and (5) models of type (1), (2), (4) with server vacations. In Chapter 1, we have given definitions of these models and a brief description of the related existing results in the queueing literature. We have introduced three major control policies: the T–policy, the N–policy and the D–policy. We have also mentioned that a systematic study of systems under these control policies does not exist in the literature, because of the difficulties involved in solving some of the systems by traditional methods. In this section, as well as in the subsequent sections, we will address this particular issue.

From the definitions of these three control policies, we see that, by interpreting the parameter T as a vacation of constant duration, the T–policy is actually a special case of the vacation model (see Doshi [1985]). As described in Chapter 1, the customer delay in the vacation model can be decomposed into the sum of two random variables. The first of the two is distributed as the customer delay in a standard M/G/1 queue and the second is distributed as the equilibrium excess (or forward recurrence time) of a typical vacation time. So, the waiting–time distribution of customers in an M/G/1 queue under the T–policy is indirectly known.

The distribution of customer delay in the N–policy case has also been studied in the

literature before (see Neuts [1981]), although the result is given in transform form. Similar to the standard M/G/1 queue, the number of customers in the system at customer departure epochs forms a Markov chain. By using standard methods for solving the M/G/1–type queues, one can derive the distribution of customer delay in the M/G/1 queue operating under the N–policy.

The distribution of customer delay for the M/G/1 queue under the D–policy does not seem to have been studied before. In this chapter, we will study the delay distribution of customers in the M/G/1 queue under D–policy by constructive sample–path approaches. The basic idea behind our analysis is to classify customers according to whether they arrive during off– or on–period of the system. If a customer arrives when the system is off, then he has to wait for the system to be turned on and for all customers in front of him to complete their services, before he can receive any service; given the number of customers present in the system when it is turned on, the waiting–time distribution for such a customer is fairly easy to compute, by conditioning on his position in queue. The main difficulty lies on analyzing the delay of customers who arrive when the system is on. What we will do in our analysis is to treat all the customers present in the system when it is turned on as a *single* "customer", whom we call a supercustomer; this supercustomer is then considered the one who initiates the busy period. With this view, the delay of customers who arrive during on–periods of the system is the same as the delay of those who do not initiate busy periods in an M/G/1 queue with "exceptional" first services in busy periods. By this consideration, we relate the analysis of M/G/1 queues under D–policy to that of M/G/1 queues with exceptional first service in busy periods.

The organization of the rest of this chapter is as follows. In Section 2.2, we analyze the delay distribution of customers in the M/G/1 queue with exceptional first services, especially of those who have "ordinary" services in this system. Because the method we use is valid also for general renewal arrival processes, we will state our results in the

context of GI/G/1 queues. We also extend our basic method of analysis to GI/G/1 queues with set–up times and GI/G/1 queues with server vacations. We show that the decom–position result for vacation models holds in more general settings (combinations of vacation models and other variations). In Section 2.3, we derive, in explicit forms, the distributions of customer delays in M/G/1 queues under the D–policy and the N–policy. In Section 2.4, we discuss some applications of the results to the optimal system control.

2.2 PRELIMINARY RESULTS AND SOME APPLICATIONS

The basic model we analyze is a GI/G/1 queue with infinite waiting space. Arriving customers are served according to the preemptive–resume–LIFO queue discipline. We assume, without loss of generality, that the first customer arrives at time 0 finding the system empty. For $i = 1, 2, \cdots$, let A_i and S_i be the arrival time and the service time of the i^{th} customer, respectively. Also, let $T_i = A_{i+1} - A_i$, $i \geq 1$; then we assume $\{T_i, i \geq 1\}$ and $\{S_i, i \geq 1\}$ are two sequences of i.i.d. (independent and identically distributed) random variables that are also independent of one another. In addition, we assume that the arrival rate $\lambda \equiv 1/E(T)$ $(0 < \lambda < \infty)$ is less than the service rate $\mu \equiv 1/E(S)$ $(0 < \mu < \infty)$, where T and S denote typical versions of interarrival and service times respectively.

We define the state of the system to be $\{j; x_1,x_2,\cdots,x_j)$, which indicates that: (1) there are j ($j \geq 1$) customers in the system; and (2) the remaining service times of these customers, arranged in increasing order of their arrival times, are respectively greater than $x_1, x_2, \cdots, x_j$. When the dimension is clear from the context, we also sometimes write $\mathbf{x}$ in place of the vector $(x_1,x_2,\cdots,x_j)$. We denote by $\alpha_j(x_1,x_2,\cdots,x_j)$ the limiting proportion of customers who, on their arrival, find the system in state $\{j; x_1,x_2,\cdots,x_j)$. Formally, $\alpha_j(x_1,x_2,\cdots,x_j)$ is defined as

$$\alpha_j(x) \equiv \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} 1_{\{j;x\}}(A_i) , \tag{2.1}$$

where $1_{\{j;x\}}(A_i)$ is the indicator function of the event that customer i finds the system in state $\{j;x\}$ on his arrival. Note that this limit converges to a constant w.p.1 (with probability 1) because of the assumption $\lambda < \mu$. In the same way, we define α_0 as the limiting proportion of customers who, on arrival, find the system empty. Note that α_0 and $\alpha_j(x_1,x_2,\cdots,x_j)$ completely describe the probabilistic behavior of workload in the system as seen by arriving customers.

Our analysis in this section focuses on the evaluation of α_0 and $\alpha_j(x_1,x_2,\cdots,x_j)$. Niu [1988] established, by a sample–average argument, that for $j \geq 1$ and $x_1,\cdots,x_j \geq 0$,

$$\frac{\alpha_j(x_1,\cdots,x_{j-1},x_j)}{\alpha_{j-1}(x_1,\cdots,x_{j-1})} = E[m_D((S-x_j)^+)] , \tag{2.2}$$

where $m_D(t)$ is the renewal function for a "delayed" renewal process (see, for example Ross [1983], p.74) whose first interevent time is distributed as T and the others as I, the idle period in a standard GI/G/1 queue; and $\alpha_{j-1}(x_1,\cdots,x_{j-1})$ is defined to be α_0 when $j = 1$. The operation $(\cdot)^+$ in the right hand–side of (2.2) is defined to be max $(0,\cdot)$. For a complete discussion, please see Niu [1988]; in the following, we give an outline of his proof for completeness.

A basic concept needed for the proof of (2.2) is that of j–cycles, $j \geq 0$. A j–cycle is defined to be a time period that begins with an arrival finding the system in state $\{j;0\}$ and ends when such an event occurs again for the next time. For a given $j \geq 1$, he evaluates the ratio of averages

$$\frac{\alpha_j(x_1,\cdots,x_{j-1},x_j)}{\alpha_{j-1}(x_1,\cdots,x_{j-1})} = \lim_{n\to\infty} \frac{\sum_{i=1}^{n} 1_{\{j;x_1,\cdots,x_{j-1},x_j\}}(A_i)/n}{\sum_{i=1}^{n} 1_{\{j-1;x_1,\cdots,x_{j-1}\}}(A_i)/n}$$

by considering successive (j–1)–cycles. Observe that $1_{\{j-1;x_1,\cdots,x_{j-1}\}}(A_{n_k}) = 0$ (where n_k is the index of the arrival epoch at which the k^{th} (j–1)–cycle begins) in a (j–1)–cycle implies that $1_{\{j;x_1,\cdots,x_{j-1},x_j\}}(A_i) = 0$ for every i in that cycle because the status of those customers who are present immediately before A_{n_k} remains unchanged as long as there are j or more customers in the system. Therefore, by ignoring (j–1)–cycles with $1_{\{j-1;x_1,\cdots,x_{j-1}\}}(A_{n_k}) = 0$, the above expression simplifies to

$$\frac{\alpha_j(x_1,\cdots,x_{j-1},\, x_j)}{\alpha_{j-1}(x_1,\cdots,\, x_{j-1})} = \lim_{n\to\infty} \frac{1}{m} \sum_{i=1}^{m} N_i(x_j \,|\, j-1;\, x_1,\cdots,x_{j-1}), \tag{2.3}$$

where m is the number of (j–1)–cycles that begin in state $\{j-1;\, x_1,\cdots,\, x_{j-1}\}$ and $N_i(x_j\,|\,j-1;x_1,\cdots,x_{j-1})$ denotes the number of arrivals who find the system in state $\{j;\, x_1,\cdots,\, x_{j-1},\, x_j\}$ during the i^{th} such (j–1)–cycles. Since such a cycle is initiated by the arrival of a "test" customer and since the queue discipline is preemptive–resume LIFO, $N_i(x_j\,|\,j-1;x_1,\cdots,x_{j-1})$ is distributed as the number of renewals in the random interval $(0,\, (S-x_j)^+)$ in a "delayed" renewal process where the first interevent time is distributed as T and the others as I. Furthermore, since *the experience of any customer is not affected in any way by those who arrive before him*, these random variables are i.i.d., independent of $\{j-1;\, x_1,\cdots,\, x_{j-1}\}$. Relation (2.2) now follows by applying the strong law of large numbers to the right–hand side of (2.3).

Applying relation (2.2) iteratively, we obtain the following explicit expression for $\alpha_j(x_1,x_2,\cdots,x_j)$:

$$\alpha_j(x) = \alpha_0[E(m_D(S))]^j \prod_{i=1}^{j} \frac{E[m_D((S_i - x_i)^+)]}{E[m_D(S)]}, \qquad \text{for } j \geq 1 \text{ and } x \geq 0; \tag{2.4a}$$

and α_0 can be derived by the normalization condition $\alpha_0 + \sum_{j=1}^{\infty} \alpha_j(0) = 1$, leading to

$$\alpha_0 = 1 - E[m_D(S)]. \tag{2.4b}$$

We now consider a GI/G/1 queue where the first customer in each busy period requires an exceptional service. Let S^1 be a typical exceptional service. For this system, we denote by α_0' and $\alpha_j'(x_1,x_2,\cdots,x_j)$ the counterparts of α_0 and $\alpha_j(x_1,x_2,\cdots,x_j)$ in the standard GI/G/1 queue. Note that relation (2.2) remains valid for α_0' and $\alpha_j'(x_1,x_2,\cdots,x_j)$ for all $j \geq 2$; and when $j = 1$, we need to replace S by S^1 in the right–hand side of (2.2). Applying relation (2.2) iteratively and using the normalization condition $\sum_{j=1}^{\infty} \alpha_j'(0) + \alpha_0' = 1$ now lead to the following theorem:

Theorem 1. Consider a GI/G/1 queue with exceptional first services in busy periods. Then, in the preemptive–resume–LIFO queue discipline, we have, for $j \geq 1$ and $x \geq 0$,

$$\alpha_j'(x) = \alpha_0' \, [E(m_D(S^1))] \, [E(m_D(S))]^{j-1} \frac{E[m_D((S^1 - x_1)^+)]}{E[m_D(S^1)]} \prod_{i=2}^{j} \frac{E[m_D((S_i - x_i)^+)]}{E[m_D(S)]}; \tag{2.5a}$$

and

$$\alpha_0' = \frac{1 - E[m_D(S)]}{1 + E[m_D(S^1)] - E[m_D(S)]}. \tag{2.5b}$$

The Laplace–Stieltjes transform of the delay distribution for queues with exceptional first services is known previously for the Poisson–arrival case (see Welch [1964], Avi–Itzhak, Maxwell and Miller [1965]). Our Theorem 1 is stated in the more general GI/G/1 setting and is proved constructively.

Formula (2.5) describes the explicit average state behavior over all arrival epochs. As mentioned in Section 2.1, to derive the delay distribution of customers in systems under D–policy, we need the average state behavior over arrival epochs of only *those who do not initiate busy periods*; and this motivates the next theorem.

Theorem 2. In a GI/G/1 queue with exceptional first services in busy periods, the distribution of workload as seen by arriving customers who do not initiate busy periods is given, for $x \geq 0$, by

$$P(W \leq x) = F * H(x) , \tag{2.6}$$

where $F(x)$ is given by $F(x) = 1 - E[m_D((S^1-x)^+)]/E[m_D(S^1)]$ and $H(x)$ is the workload distribution at the arrival epoch of a randomly selected customer in a corresponding standard GI/G/1 queue, given by

$$H(x) = \alpha_0 + \sum_{j=1}^{\infty} \alpha_j(\mathbf{0})\Psi^{[j]}(x) , \tag{2.7}$$

where $$\Psi(x) = \left[1 - \frac{E[m_D((S-x)^+)]}{E[m_D(S)]}\right],$$

and [j] stands for j–fold convolution operation.

Proof. Equation (2.5) gives the complete description of workload behavior at the arrival epochs of *all* customers, including those who initiate busy periods. In order to describe the workload behavior at arrival epochs of only those who do not initiate busy periods, we exclude from consideration customers who find the system empty on their arrival by working with "relative proportions" $\alpha''_j(x_1,x_2,\ldots,x_j)$ for $j \geq 1$ and $\mathbf{x} \geq \mathbf{0}$, defined by

$$\alpha''_j(x_1,x_2,\ldots,x_j) = \frac{1}{1 - \alpha'_0}\, \alpha'_j(x_1,x_2,\ldots,x_j) .$$

Substitution of (2.5) into the right–hand side of the above expression yields

$$\alpha''_j(x_1,x_2,\ldots,x_j) = \frac{E[m_D((S^1-x_1)^+)]}{E[m_D(S^1)]}\{1-E[m_D(S)]\}\prod_{i=2}^{j}\frac{E[m_D((S_i-x_i)^+)]}{E[m_D(S)]}$$

$$= \frac{E[m_D((S^1-x_1)^+)]}{E[m_D(S^1)]}\alpha_{j-1}(x_2,\ldots,x_j), \qquad (2.8)$$

where the second equality is due to (2.4a). Since the right–hand side of (2.8) is the *product* of $\bar{F}(x_1)$ and $\alpha_{j-1}(x_2,\ldots,x_j)$ and since $\alpha_{j-1}(x_2,\ldots,x_j)$ for $j \geq 1$ completely describe the distribution of the workload as seen by arrivals in a standard GI/G/1 queue, our proof is completed.

The argument used in the proof of Theorem 2 is very useful. By properly introducing an "artificial" customer at the beginning of each busy period (this artificial customer therefore initiates a busy period), we can derive the probability distribution of workload as seen by arrivals in a variety of related models. As examples, we will use this method to analyze two specific models: GI/G/1 queues with set–up times and GI/G/1 queues with server vacations.

Example 1. GI/G/1 Queue with Set–up Times

In this model, at the beginning of each busy period, the server spends S^1, with distribution of $G_s(\cdot)$, units of set–up time before starting to serve customers. Conceptually, this is equivalent to saying that at the beginning of each busy period, the server sequentially serves *two* customers: an "artificial" customer with service time S^1 followed immediately by a "genuine" customer with an ordinary service time. The following theorem is a direct consequence of this interpretation.

Theorem 3. In a GI/G/1 queue with set–up times, the distribution of the workload in the system as seen by arriving customers is given by

$$P(W \le x) = F_s^* H(x), \tag{2.9}$$

where

$$F_s(x) = 1 - \overline{G}_s(x)\,\frac{1+E[M(S^1-x)\,|\,S^1 \ge x]}{1+E[M(S^1)]},$$

where $M(\cdot)$ is the renewal function of a renewal process whose interevent time is distributed as the idle period in a standard GI/G/1 queue, and $H(x)$ is as given in (2.7).

Proof. Since in this system the "artificial" customer is followed immediately by a genuine arrival, the first interevent time in the "delayed" renewal process described in the proof of (2.2) is no longer distributed as T; it is instead a constant time interval with duration zero. Therefore, when $j = 1$, the term $E[N_i(x_j\,|\,j-1;\ x_1, \cdots, x_{j-1})]$ in (2.3) simplifies to $\bar{G}(x_1) + E[M((S^1-x_1)^+)]$, where $\bar{G}(x_1)$ is the contribution to the total count by the genuine customer who arrives immediately after the arrival of the artificial customer. The rest of the argument follows the same line as in the proof of Theorem 2.

Intuitively, if we ignore the part of the workload contributed by the "artificial" customer, the rest of the workload in the system (contributed by "genuine" customers) behaves exactly the same way as the workload in a standard GI/G/1 queue, because the artificial customer has no effect whatsoever on the behavior of genuine customers when the queue discipline is preemptive–resume LIFO. Thus, the only unknown is the steady state behavior of the remaining "artificial service" at the arrival epochs of genuine customers. Theorem 3 actually gives the explicit expression of the distribution of this "remaining service". To verify this expression, we specialize S^1 to an exponentially distributed random variable. Thus, the remaining workload contributed by the "artificial" customer and

observed by genuine arrivals, due to the memoryless property of the exponential distribution, is exponentially distributed. In other words, when the set–up time is exponential, the total workload in the system as seen by genuine customers is distributed as the sum of S^1 and the workload as seen by arrivals in a standard GI/G/1 queue. As a formal check, we have, by the memoryless property, that

$$E[M(S^1-x)\,|\,S^1>x] = E[M(S^1-x)],$$

and therefore $F_s(x) = 1 - \bar{G}_s(x) = G_s(x)$. We see, (2.9) indeed simplifies as expected. Theorem 3 has been obtained in the literature before, in Laplace–Stieltjes–transform form (see Doshi [1985]). Our result is explicit and our argument slightly more constructive.

Example 2. GI/G/1 Queue with Server Vacation

The definition of the vacation model was given in Chapter 1. Clearly, this model differs from the previous model only in the starting condition at the beginning of each busy period. By treating each vacation as an "artificial" service, we can also solve this model by the method described in the previous example. In the following analysis, we assume that the vacation times V_i, $i = 1, 2, \cdots$, are i.i.d. random variables.

Theorem 4. Under the condition that at least one of the three random variables T, S, V is nonlattice, the workload as seen by arriving customers in a GI/G/1 queue with server vacation is distributed as the sum of two random variables; one of them is distributed as the workload as seen by arriving customers in a standard GI/G/1 queue, and the other the equilibrium excess (or forward recurrence time) of vacation.

Proof. The decomposition result of Theorem 4 can be easily established by treating the vacation times as "artificial" service times and using the same reasoning as we did in the previous example. The only unknown is the distribution of the remaining "artificial"

service in the system as seen by "genuine" arrivals. In the preemptive–resume–LIFO queue discipline, we give, in figure 1, a typical realization of the contribution to the total workload from the vacations (that is, after deleting all workload brought in by genuine customers).

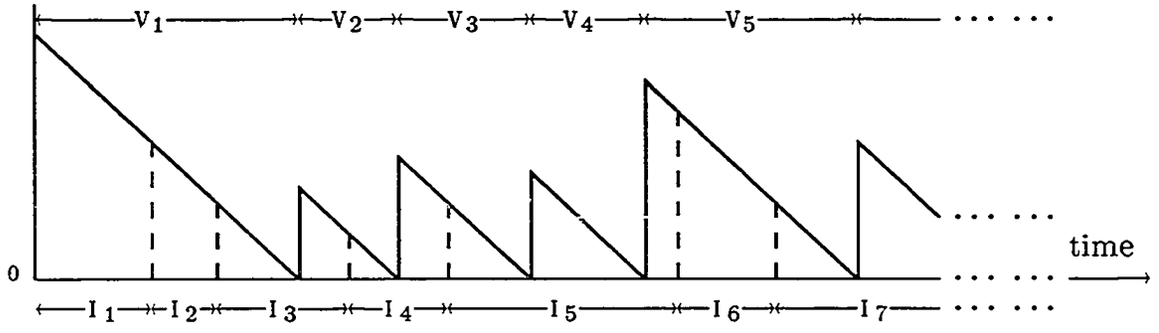

Figure 1. Contracted version of a typical realization of the "vacation workload" in a GI/G/1 queue with server vacation, in the preemptive–resume–LIFO queue discipline.

It is easy to see from figure 1 that the successive vacations and the successive idle periods are two independent renewal processes but both start at the same time (t = 0). Therefore, when t goes to infinity, observed from a renewal point of the idle–period process, the time until the completion of the next vacation is distributed as the equilibrium excess (or forward recurrence time) of a vacation. The proof is completed by noting that all genuine customers in a "genuine busy period" see the same remaining vacation time.

It is important to notice that *this proof holds for many other infinite–capacity queueing models that have server vacations*, such as vacation models with exceptional first service, with set–up times, or under certain control policies. Figure 1 forms the basis for the analyses of all these models. The only necessary modification is that the successive idle periods for different variations of the vacation models are different. In fact, Lee and Srinivasan [1989] have discovered that the mean waiting time of customers in an M/G/1 vacation model under the N–policy is decomposed into two parts: the mean residual

vacation time and the mean waiting time in an M/G/1 queue under the N–policy. They also commented that their result might be derived by applying vacation model results while treating the M/G/1 queue under N–policy as a standard M/G/1 queue, but they were unable to justify this approach formally. We have shown here that such decomposition results exist for more general queueing systems that combine server vacations with other features, as described above.

2.3 WAITING–TIME DISTRIBUTIONS IN M/G/1 QUEUES UNDER D–POLICY AND N–POLICY

In the previous section, we analyzed the average workload behavior over arrival epochs of customers in several related GI/G/1 models. All of the models analyzed in the previous section have modified starting conditions at the beginning of busy periods. These modifications were reinterpreted as "services" brought in by carefully defined "artificial" customers. Because all the services brought in by these "artificial" customers are independent of the arrival processes and the service processes in these systems, such reinterpretations greatly simplify our analysis. But, for the M/G/1 queues under D–policy and N–policy, we will no longer benefit from such convenient and direct reinterpretations, because the "artificial services" in these systems are dependent of the arrival process and the service process. Therefore, we have to modify our methods of analysis to avoid this "dependence" and try to use indirectly the results obtained in the previous section.

In the M/G/1 queue under D–policy, customers are served according to the non–preemptive FIFO queue discipline. We classify customers into two types; those who arrive during off–periods of the system are called type one, and those during on–periods type two. The busy period in this model is defined as the time interval that starts at the arrival epoch of a customer who finds the system empty and ends at the departure epoch of

a customer who leaves the system empty. Furthermore, let an n–busy period be one that has exactly n type–one customers; note that a busy period will be an n–busy period with probability b_n, given by

$$b_n = P(\sum_{i=1}^{n-1} S_i \le D, \sum_{i=1}^{n} S_i > D) = G^{[n-1]}(D) - G^{[n]}(D) , \tag{2.10}$$

where G(x) is the service–time distribution. In other words, b_n is the probability that the work brought in by the n^{th} arrival makes the total workload in the system exceed the threshold D.

Now, suppose that a customer (called the "test" customer) is selected randomly from the infinite poll of all customers. Denote by A_n, $n \ge 1$, the event that this customer arrives in an n–busy period. Given that our test customer arrives in an n–busy period, the distribution of his delay depends further on whether he is a type–one customer or a type–two. Suppose that he is a type–one customer and is in the k^{th} position of the queue before the system is turned on; then he has to wait for the n^{th} customer to arrive (the system is turned on then) and for all the k–1 customers in front of him to complete their services before he can receive any service. The time until the arrival of the n^{th} customer is simply the sum of n–k exponentially distributed random variables with rate λ; this distribution will be denoted by $E_{n-k}(x)$ ($E_0(x) \equiv 1$). Denote by $V_{kn}(x)$ the distribution of the total service time requested by the k–1 customers in front of our test customer; then

$$V_{1n}(x) = 1, \tag{2.11a}$$

and, for $k = 2, \cdots, n$,

$$V_{kn}(x) = P(\sum_{i=1}^{k-1} S_i \le x \;\Big|\; \sum_{i=1}^{n-1} S_i \le D, \sum_{i=1}^{n} S_i > D)$$

$$= \frac{P(\sum_{i=1}^{k-1} S_i \le x,\ \sum_{i=1}^{n-1} S_i \le D,\ \sum_{i=1}^{n} S_i > D)}{P(\sum_{i=1}^{n-1} S_i \le D,\ \sum_{i=1}^{n} S_i > D)}.$$

By conditioning on the values of $\sum_{i=1}^{k-1} S$, we obtain, for $k = 2, \cdots, n$ that

$$V_{kn}(x) = \frac{1}{b_n} \int_0^x P\left[\sum_{i=1}^{n-1} S_i \le D,\ \sum_{i=1}^{n} S_i > D \ \middle|\ \sum_{i=1}^{k-1} S_i = t\right] dG^{[k-1]}(t)$$

$$= \frac{1}{b_n} \int_0^x \left[\bar{G}^{[n-k+1]}(D-t) - \bar{G}^{[n-k]}(D-t)\right] dG^{[k-1]}(t). \qquad (2.11b)$$

Since the test customer is equally likely to be at any one of the n positions, given that he is a type–one customer in an n–busy period, his delay distribution is given by

$$\frac{1}{n} \sum_{k=1}^{n} V_{kn} * E_{n-k}(x). \qquad (2.12)$$

Now, if the test customer is type–two, then his delay distribution can be obtained from (2.6) (Theorem 2) by specializing it to the Poisson–arrival case; the idea is to treat all type–one customers in an n–busy period as a single "super–customer". Because the arrival of this super–customer initiates an n–busy period, the delay distribution of type–two customers in an n–busy period obviously follows the distribution (2.6). Denote by U_n the total service time of the super customer in an *n–busy period*, then F(x) in (2.6) becomes

$$F_n(x) = 1 - \frac{E[(U_n - x)^+]}{E(U_n)}, \qquad (2.13)$$

where we have used the relation $m_D(t) = \lambda t$ for the Poisson arrival and added a subscript n to F to indicate that it is associated with an *n–busy period.* Since $U_n = \sum_{i=1}^{n} S_i$ and since $\sum_{i=1}^{n-1} S_i \le D$ and $\sum_{i=1}^{n} S_i > D$ in any n–busy period, we have

$$P(U_1 \le x) = 1 - \bar{G}(x)/\bar{G}(D) \tag{2.14a}$$

and, for $n > 1$, we have

$$P(U_n \le x) = P(U_n \le x \mid \sum_{i=1}^{n-1} S_i \le D, \sum_{i=1}^{n} S_i > D)$$

$$= \frac{P(\sum_{i=1}^{n} S_i \le x, \sum_{i=1}^{n-1} S_i \le D, \sum_{i=1}^{n} S_i > D)}{P(\sum_{i=1}^{n-1} S_i \le D, \sum_{i=1}^{n} S_i > D)}.$$

By conditioning on the values of $\sum_{i=1}^{n-1} S_i$, the above expression can be simplified as follows:

$$P(U_n \le x) = \frac{1}{b_n} \int_0^D P(D < \sum_{i=1}^{n} S_i \le x \mid \sum_{i=1}^{n-1} S_i = t)\, dG^{[n-1]}(t)$$

$$= \frac{1}{b_n} \int_0^D [G(x-t) - G(D-t)]\, dG^{[n-1]}(t)$$

$$= \frac{1}{b_n} \left[\int_0^D G(x-t)\, dG^{[n-1]}(t) - G^{[n]}(D) \right]. \tag{2.14b}$$

We also have

$$E(U_1) = \int_0^\infty P(U_1 > x)\, dx = D + \int_D^\infty P(U_1 > x) dx$$

$$= D + [\mu - \int_0^\infty \overline{G}(x)\, dx]/\overline{G}(D), \tag{2.15a}$$

where $\mu = E(S)$; and, for $n > 1$,

$$E(U_n) = \int_0^\infty P(U_n > x) dx = D + \int_D^\infty P(U_n > x)\, dx$$

$$= D + \int_D^\infty \left[1 - \frac{1}{b_n} \int_0^D [G(x-t) - G(D-t)]\, dG^{[n-1]}(t) \right] dx$$

$$= D + \frac{1}{b_n}\int_D^\infty \left[G^{[n-1]}(D) - G^{[n]}(D) - \int_0^D G(x-t)\, dG^{[n-1]}(t) + G^{[n]}(D) \right] dx$$

$$= D + \frac{1}{b_n}\int_D^\infty \left[\int_0^D dG^{[n-1]}(t) - \int_0^D \overline{G}(x-t)\, dG^{[n-1]}(t) \right] dx$$

$$= D + \frac{1}{b_n}\int_D^\infty \int_0^D \overline{G}(x-t)\, dG^{[n-1]}(t)\, dx.$$

By exchanging the order of integration in the last equality, we obtain

$$E(U_n) = D + \frac{1}{b_n}\int_0^D \int_D^\infty \overline{G}(x-t)\, dx\, dG^{[n-1]}(t)$$

$$= D + \frac{1}{b_n}\int_0^D \int_{D-t}^\infty \overline{G}(u)\, du\, dG^{[n-1]}(t).$$

Exchanging the order of integrations one more time, we obtain

$$E(U_n) = D + \frac{1}{b_n}\left[\mu G^{[n-1]}(D) - \int_0^D \overline{G}(u) G^{[n-1]}(D-u)\, du \right]. \qquad (2.15b)$$

When a specific form of service distribution is given, (2.15b) is a formula ready for computation. We have therefore determined $F_n(x)$ completely.

We now compute the probability that the test customer is of type–one, given that he arrives in an n–busy period. Let K_n to be the number of customers served in an n–busy period. Clearly, there must be n type–one customers in an n–busy period; and during the service period of these n customers, the average number of type two customers arrived to the system is $\lambda E(U_n)$, and during the service period of these type–two customers, the average number of the newly arrived type–two customers is $\lambda[\lambda E(U_n)\mu]$ and so forth. This reasoning leads to

$$E(K_n) = n + \lambda E(U_n) + \lambda E(U_n)\rho + \lambda E(U_n)\rho^2 + \cdots \cdots$$

$$= n + \lambda E(U_n)/(1-\rho). \tag{2.16}$$

Since the first term in the right–hand side of (2.16) is the number of type–one customers in an n–busy period and the second term the mean number of type–two customers who arrived in an n–busy period, it follows that the probability for a randomly selected customer to be of type one is

$$\frac{n}{E(K_n)} = \frac{n(1-\rho)}{n(1-\rho)+\lambda E(U_n)}, \tag{2.17a}$$

and that of type two is

$$1 - \frac{n}{E(K_n)} = \frac{\lambda E(U_n)}{n(1-\rho)+\lambda E(U_n)}. \tag{2.17b}$$

Combining (2.12), (2.6), and (2.17) now leads to the delay distribution for customers arriving within n–busy periods:

$$P(W \le x \mid A_n) = \frac{n}{E(K_n)}\left[\frac{1}{n}\sum_{k=1}^{n} V_{kn}{}^{*}E_{n-k}(x)\right] + \left[1 - \frac{n}{E(K_n)}\right]F_n{}^{*}H(x)$$

$$= \frac{1-\rho}{n(1-\rho)+\lambda E(U_n)}\sum_{k=1}^{n} V_{kn}{}^{*}E_{n-k}(x) + \frac{\lambda E(U_n)}{n(1-\rho)+\lambda E(U_n)}F_n{}^{*}H(x), \tag{2.18}$$

where H(x) is the delay distribution of a randomly selected customer in a standard M/G/1 queue. By "unconditioning", we obtain the delay distribution for a randomly selected customer in the M/G/1 queue under D–policy:

$$P(W \le x) = \sum_{n=1}^{\infty} a_n\left\{\frac{1-\rho}{n(1-\rho)+\lambda E(U_n)}\sum_{k=1}^{n} V_{kn}{}^{*}E_{n-k}(x) + \frac{\lambda E(U_n)}{n(1-\rho)+\lambda E(U_n)}F_n{}^{*}H(x)\right\}, \tag{2.19}$$

where $a_n = P(A_n)$, the probability for a randomly selected customer to arrive in an n–busy period. Note that a_n does not equal to b_n (see (2.10)), because the average number of customers served in different types of busy periods is different, leading to length–biasing. To correct this bias, we interpret the number of customers served in a busy period as

"sojourns" of a discrete–time semi–Markov process, and apply Theorem 8.3 of Ross [1983] to obtain

$$a_n = \frac{E(K_n)b_n}{\Sigma_n E(K_n)b_n} = \frac{E(K_n)\,b_n}{E(K)} = \frac{b_n}{E(K)}\left[\frac{n(1-\rho)+\lambda E(U_n)}{1-\rho}\right], \tag{2.20}$$

where, K denotes the number of customers served in a randomly selected busy period. Substitution of (2.20) into (2.19) now leads to the following:

Theorem 5. The delay distribution of a randomly selected customer in an M/G/1 queue under D–policy is given by

$$P(W \le x) = \sum_{n=1}^{\infty} \frac{b_n}{E(K)} \left\{ \sum_{k=1}^{n} V_{kn} * E_{n-k}(x) + \frac{\lambda E(U_n)}{1-\rho} F_n * H(x) \right\}. \tag{2.21}$$

Theorem 6. The expected delay of customers in the system described in Theorem 5 is given by

$$E(W) = \frac{1}{1+m(D)}\left[\int_0^D t\, dm(t) + \frac{1-\rho}{\lambda}\int_0^D m(D-t)\, dm(t) + \frac{1-\rho}{\lambda} m(D)\right] + E(W_M). \tag{2.22}$$

where $m(t)$ is the renewal function defined as $m(t) = \sum_{n=0}^{\infty} G^{[n]}(D)$ and $E(W_M)$ is the expected delay of a randomly selected customer in a standard M/G/1 queue.

Proof. By definition, $E(W) = \int_0^{\infty} t\, dP(W \le t)$, where $P(W \le t)$ is given by (2.21). Applying (2.21) and changing the order of integration and summation, we obtain

$$E(W) = \int_0^{\infty} t\, d\left\{ \sum_{n=1}^{\infty} \frac{b_n}{E(K)} \left[\sum_{k=1}^{n} V_{kn} * E_{n-k}(t) + \frac{\lambda E(U_n)}{1-\rho} F_n * H(t) \right] \right\}$$

$$= \sum_{n=1}^{\infty} \frac{b_n}{E(K)} \left\{ \int_0^{\infty} t\, d\left[\sum_{k=1}^{n} V_{kn} * E_{n-k}(t) \right] + \frac{\lambda E(U_n)}{1-\rho} \int_0^{\infty} t\, dF_n * H(t) \right\}$$

$$= \sum_{n=1}^{\infty} \frac{b_n}{E(K)} \left\{ \sum_{k=1}^{n} \int_0^{\infty} t\, d\left[V_{kn} * E_{n-k}(t) \right] + \frac{\lambda E(U_n)}{1-\rho} \int_0^{\infty} t\, d[F_n * H(t)] \right\}.$$

When the distribution function of a random variable is expressed in convolution form, the expectation of this random variable can be written as sum of the expectations of the component random variables. That is,

$$E(W) = \sum_{n=1}^{\infty} \frac{b_n}{E(K)} \left\{ \sum_{k=1}^{n} \left[\int_0^{\infty} t \, dV_{kn}(t) + \int_0^{\infty} t \, dE_{n-k}(t) \right] + \frac{\lambda E(U_n)}{1-\rho} \left[\int_0^{\infty} t \, dF_n(t) + \int_0^{\infty} t \, dH(t) \right] \right\}.$$

Noting that $F_n(t)$ is a equilibrium distribution, we have

$$E(W) = \sum_{n=1}^{\infty} \frac{b_n}{E(K)} \left\{ \sum_{k=1}^{n} \left[\int_0^{\infty} t \, dV_{kn}(t) + (n-k)\frac{1}{\lambda} \right] + \frac{\lambda E(U_n)}{1-\rho} \left[\frac{E(U_n^2)}{2E(U_n)} + E(W_M) \right] \right\}$$

$$= \sum_{n=1}^{\infty} \frac{b_n}{E(K)} \left\{ \sum_{k=1}^{n} \left[\int_0^{\infty} t \, dV_{kn}(t) + (n-k)\frac{1}{\lambda} \right] + \frac{\lambda E(U_n^2)}{2(1-\rho)} + \frac{\lambda E(U_n)}{1-\rho} E(W_M) \right\}.$$

Substituting $V_{kn}(t)$ in (2.11) to the above equation, we obtain

$$E(W) = \frac{1}{E(K)} \left\{ \sum_{n=2}^{\infty} \sum_{k=2}^{n} \int_0^D t \left[G^{[n-k]}(D-t) - G^{[n-k+1]}(D-t) \right] dG^{[k-1]}(t) + \right.$$

$$\left. + \sum_{n=1}^{\infty} b_n \frac{n(n-1)}{2\lambda} + \sum_{n=1}^{\infty} b_n \frac{\lambda E(U_n^2)}{2(1-\rho)} + \sum_{n=1}^{\infty} b_n \frac{\lambda E(U_n)}{1-\rho} E(W_M) \right\}$$

$$= \frac{1}{E(K)} \left\{ \sum_{n=2}^{\infty} \sum_{k=2}^{n} \int_0^D t \left[G^{[n-k]}(D-t) - G^{[n-k+1]}(D-t) \right] dG^{[k-1]}(t) + \right.$$

$$\left. + \frac{1}{2\lambda} \sum_{n=1}^{\infty} \left[G^{[n-1]}(D) - G^{[n]}(D) \right] (n^2+n) + \frac{\lambda E(U^2)}{2(1-\rho)} + \frac{\lambda E(U)}{1-\rho} E(W_M) \right\},$$

where $E(U) = \sum_{n=1}^{\infty} b_n E(U_n)$ and $E(U^2) = \sum_{n=1}^{\infty} b_n E(U_n^2)$. By changing the order of summations and the order of integration and summation in the above equation, we obtain

$$E(W) = \frac{1}{E(K)} \left\{ \sum_{k=2}^{\infty} \sum_{n=k}^{\infty} \int_0^D t \left[G^{[n-k]}(D-t) - G^{[n-k+1]}(D-t) \right] dG^{[k-1]}(t) + \right.$$

$$\left. + \frac{1}{2\lambda} \left[\sum_{n=1}^{\infty} n^2 G^{[n-1]}(D) - \sum_{n=1}^{\infty} n^2 G^{[n]}(D) - \sum_{n=1}^{\infty} n G^{[n-1]}(D) + \sum_{n=1}^{\infty} n G^{[n]}(D) \right] + \right.$$

$$+ \frac{\lambda E(U^2)}{2(1-\rho)} + \frac{\lambda E(U)}{1-\rho} E(W_M) \Big\}.$$

$$= \frac{1}{E(K)} \Big\{ \sum_{k=2}^{\infty} \int_0^D t \Big[\sum_{n=k}^{\infty} G^{[n-k]}(D-t) - \sum_{n=k}^{\infty} G^{[n-k+1]}(D-t) \Big] \, dG^{[k-1]}(t) +$$

$$+ \frac{1}{2\lambda} \Big[1 + \sum_{n=1}^{\infty} (n+1)^2 G^{[n]}(D) - \sum_{n=1}^{\infty} n^2 G^{[n]}(D) - 1 - \sum_{n=1}^{\infty} (n+1) G^{[n]}(D) +$$

$$+ \sum_{n=1}^{\infty} n G^{[n]}(D) \Big] + \frac{\lambda E(U^2)}{2(1-\rho)} + \frac{\lambda E(U)}{1-\rho} E(W_M) \Big\}.$$

$$= \frac{1}{E(K)} \Big\{ \int_0^D t \, d \sum_{k=2}^{\infty} G^{[k-1]}(t) + \frac{1}{\lambda} \sum_{n=1}^{\infty} n G^{[n]}(D)) + \frac{\lambda E(U^2)}{2(1-\rho)} + \frac{\lambda E(U)}{1-\rho} E(W_M) \Big\}.$$

$$= \frac{1}{E(K)} \Big\{ \int_0^D t \, dm(t) + \frac{1}{\lambda} \sum_{n=1}^{\infty} n G^{[n]}(D)) + \frac{\lambda E(U^2)}{2(1-\rho)} + \frac{\lambda E(U)}{1-\rho} E(W_M) \Big\}. \tag{2.23}$$

To simplify (2.23), we use the relations

$$E(K) = \frac{1+m(D)}{1-\rho} \tag{2.24a}$$

$$E(U) = \mu \, [1 + m(D)] \tag{2.24b}$$

$$E(U^2) = E(S^2)[1 + m(D)] + 2\mu \int_0^D t \, dm(t) \tag{2.24c}$$

(see Balachandran and Tijms [1975]) and

$$\sum_{n=1}^{\infty} n G^{[n]}(D) = \int_0^D m(D-t) \, dm(t) + m(D), \tag{2.24d}$$

(see Feller [1968] p.386) into (2.23) to obtain (2.22). The proof is completed.

We now analyze the distribution and the mean of customers delay in the M/G/1 queue under the N–policy. To study this system, we again classify customers into types one and two in the same way as we did in the D–policy case. But, for the N–policy case, we only have one type of n–busy period (and therefore no length biasing), and the total

service time of the super–customer is simply the sum of N i.i.d. regular service times (and therefore a simpler expression of the service time of our "supercustomer"), making the analysis of distribution of customer delay straight forward. The results are summarized in the following theorem.

Theorem 7. The waiting–time distribution of a randomly selected customer in an M/G/1 queue under N–policy is given by

$$P(W \le x) = \frac{1-\rho}{N} \sum_{k=1}^{N} G^{[k-1]} * E_{N-k}(x) + \rho\, F^* H(x), \tag{2.25}$$

where $F(x) = 1 - E[(\sum_{i=1}^{N} S_i - x)^+]/(N\mu)$; and the expected delay is given by

$$E(W) = \frac{N-1}{\lambda} + E(W_M). \tag{2.26}$$

The explicit formula (2.25) is new, and is in agreement with a transform formula given by Neuts [1981]. Also (2.26) agrees with a result given by Yadin and Naor [1963].

2.4 COMPARISONS BETWEEN OPTIMAL CONTROL POLICIES

As an application of the results obtained in the previous section, we will, in this section, compare M/G/1 queues under D–policy, N–policy and T–policy with respect to their long–run average operating costs.

Generally speaking, there are three natural cost components for the operation of a queueing system: (1) the running cost, that is, whenever the system is on, there is a linear cost of c_0 dollars per unit time; (2) the switching cost, which is a cost associated with switching the system on and off, at respective cost c_{11} and c_{12} dollars; and (3) the waiting cost, which is assumed to be an increasing function of customer delay.

Of these three types of costs, the long–run average running cost is given by a

constant $c_0\rho$, regardless of the type of policies being used and of the parameters values. To explain, we consider a stable GI/G/1 queue and define the service station alone as our "system". From $L = \lambda W$ we have that the proportion of time the server is busy equals to ρ, regardless of what type of policy is being used. So, for comparison purpose, from now on, we will exclude this cost from consideration.

The switching cost in a busy cycle is also a constant, which can be written as $c_i = c_{11} + c_{12}$, because there are exactly one switching–on and one switching–off in each busy cycle.

The waiting cost is probably the most complicated cost component among the three, because it is generally an intangible cost due to customers' negative impression on the system. Naturally, it is not easy to determine the specific form of this cost component. But in general, the analyses of this cost component requires the distribution of customer delay in the system. Note that the distribution for customer delay in the M/G/1 queue under the D–policy is not available before.

In the literature, there exist two different cost assumptions for the M/G/1 queues under control policies. Yadin and Naor [1963] proposed a cost structure in which the total cost is the sum of the second and the third cost components described above, where the third cost component is assumed to be a linear function of *customer delay.* Balachandran [1973] proposed another cost structure which differs from the one by Yadin and Naor in the third component. In his cost structure, this third component is replaced by a cost that is assumed to be a linear function of *time–average* workload in the system. Balachandran and Tijms [1975] conjectured that with respect to this cost structure, the D–policy is superior to the N–policy. This conjecture was later proved by Boxma [1976]. In reality, we do not see many examples of systems in which the intangible costs are purely due to the existence of workload instead of due to the customer delay. Therefore, there is a practical

need to compare these different control policies based on the cost structure proposed by Yadin and Naor [1963]. It is also of interest to know if the conjecture proposed by Balachandran and Tijms [1975] and prcved by Boxma [1976] still holds under this different cost structure.

In the M/G/1 queue under the D–policy, the long–run average cost (per customer) in the cost structure proposed by Yadin and Naor [1963] is given by

$$C(D) = \frac{c_1}{E(K)} + c_2 E(W), \tag{2.27}$$

where K is as defined in the previous section and E(W) is given by (2.22). Substituting (2.24a), (2.22) into (2.28), we obtain

$$C(D) = \frac{c_2}{1+m(D)}\left[\frac{c_1}{c_2}(1-\rho) + \int_0^D [t + \frac{1-\rho}{\lambda} m(D-t)]dm(t) + \frac{1-\rho}{\lambda} m(D)\right] + c_2 E(W_M). \tag{2.28}$$

In general, it is difficult to minimize this function without a specific service–time distribution, because the renewal function m(t) in (2.28) is unknown. In what follows, we consider the special case of exponentially distributed service times. In this case, $m(t) = t/\mu$ and (2.28) simplifies to

$$\begin{aligned} C(D) &= \frac{\mu c_2}{\mu+D}\left[\frac{c_1}{c_2}(1-\rho) + \frac{D^2}{2\mu} + \frac{1-\rho}{\rho}\frac{D^2}{2\mu} + \frac{1-\rho}{\rho} D\right] + c_2 E(W_M) \\ &= \frac{\mu c_2}{\mu+D}\left[\frac{c_1}{c_2}(1-\rho) + \frac{1}{\rho}\frac{D^2}{2\mu} + \frac{1-\rho}{\rho} D\right] + c_2 E(W_M). \end{aligned} \tag{2.29}$$

Since (2.29) is a differentiable convex function and $D \geq 0$, there exists a point of D, denoted by D*, such that $C(D^*) = \min_{D<\infty} (D)$. Suppose D* satisfies the condition: $\frac{d}{dD}C(D)\Big|_{D^*} = 0$. By taking the derivative of (2.29) with respect to D and using (2.29), we have

$$\frac{d}{dD}C(D) = -\frac{1}{\mu+D} C(D) + \frac{\mu c_2}{\mu+D}\left[\frac{D}{\rho\mu} + \frac{1-\rho}{\rho}\right] + \frac{c_2}{\mu+D} E(W_M) . \tag{2.30}$$

Letting (2.30) be zero leads to

$$C(D^*) = \frac{c_2}{\rho}[D^* + (1-\rho)\mu] + c_2 E(W_M). \tag{2.31}$$

The optimal value D^* can be determined by (2.29) and (2.31), i.e.

$$\frac{\mu c_2}{\mu + D^*}\left[\frac{c_1}{c_2}(1-\rho) + \frac{1}{\rho}\frac{D^{*2}}{2\mu} + \frac{1-\rho}{\rho}D^*\right] = \frac{c_2}{\rho}[D^* + (1-\rho)\mu],$$

or equivalently,

$$D^{*2} + 2\mu D^* + 2(1-\rho)\mu^2 - 2\rho(1-\rho)\mu\frac{c_1}{c_2} = 0,$$

which leads to,

$$D^* = \mu\sqrt{1 + 2(1-\rho)(\lambda\frac{c_1}{c_2} - 1)} - \mu. \tag{2.32}$$

For D^* to be real, the expression inside the square root of (2.32) must be positive, which gives the condition

$$\frac{c_1}{c_2} \geq \frac{1}{\lambda}\left[1 - \frac{1}{2(1-\rho)}\right]. \tag{2.33a}$$

If $D^* > 0$, then we must have

$$\frac{c_1}{c_2} > \frac{1}{\lambda}. \tag{2.33b}$$

Comparison between (2.33a) and (2.33b) shows that the minimum cost is achieved at a positive D^* if and only if condition (2.33b) is satisfied, because condition (2.33b) includes condition (2.33a).

In the case that (2.33b) is not satisfied, the mimimum cost will be achieved at the boundary, namely $D^* = 0$. In this case, the system operates as a standard M/G/1 queue. In the following, we will only consider the case in which (2.33b) is satisfied.

For further evaluation of C(D*), we let $\frac{c_1}{c_2} = \frac{\alpha}{\lambda}$ and substitute this ratio into (2.33) and (2.32); then, the minimum cost can be expressed as

$$C(D^*) = \frac{c_2}{\lambda}\left[\sqrt{1 + 2(1-\rho)(\alpha-1)} - \rho\right] + c_2 E(W_M), \tag{2.34}$$

where $\alpha > 1$ because of (2.33b).

It is known that under the same cost structure as we used for the D–policy case, the minimum costs for M/G/1 queues operating under the T–policy and the N–policy are, respectively, given by

$$C(T^*) = \sqrt{2c_1c_2(1-\rho)/\lambda} + c_2 E(W_M) = c_1\sqrt{2(1-\rho)/\alpha} + c_2 E(W_M), \tag{2.35}$$

$$C(N^*) = \sqrt{2c_1c_2(1-\rho)/\lambda} - \frac{c_2}{2\lambda} + c_2 E(W_M) = c_1\sqrt{2(1-\rho)/\alpha} - \frac{c_1}{2\alpha} + c_2 E(W_M), \tag{2.36}$$

where T* and N* are optimal values of T and N that minimize the long–run average costs for M/G/1 queues operating under the T–policy and the N–policy respectively (see, e.g., Cooper [1981], pp 243–253). And the value N* is given by

$$N^* = \sqrt{2\alpha(1-\rho)}, \tag{2.37}$$

(which will be used later). By definition, N* is an integer, but we treat N* as a real number for convenience. We consider the case $N^* > 1$ only (because the system would be standard M/G/1 queue otherwise). From (2.34) and (2.35), we have

$$C(D^*) - C(T^*) = -\frac{\rho(1-\rho)^2 + \sqrt{2\alpha(1-\rho)}}{\sqrt{1+2(1-\rho)(\alpha-1)} + \sqrt{2\alpha(1-\rho)}}. \tag{2.38}$$

It can be easily seen from (2.38) that $C(D^*) - C(T^*) < 0$ because the right–hand side of (2.39) is negative. Therefore, the D–policy is superior to the T–policy. It is also known that the N–policy is superior to the T–policy (see, e.g., Cooper [1981], pp. 243–253).

Hence, the T–policy is the worst policy among these three policies.

Now, we compare the D–policy and the N–policy. From (2.35) and (2.36), we have

$$C(D^*) - C(N^*) = \frac{c_1}{2\alpha}(1-2\rho)\left[1 - \frac{2}{\sqrt{2\alpha(1-\rho)+2\rho-1} + \sqrt{2\alpha(1-\rho)}}\right]. \tag{2.39}$$

It is easy to see that in order for (2.39) to be negative, either $1-2\rho$ or the expression in the square brackets must be negative. By analyzing these two expressions, we conclude that D–policy is superior to N–policy if and only if $\frac{1}{2} < \rho < 2.5 + 2\sqrt{4\alpha^2-3\alpha} - 4\alpha$. Note that $\rho < 2.5 + 2\sqrt{4\alpha^2-3\alpha}$ implies $N^* \geq 1$, a condition automatically satisfied given that $\rho > \frac{1}{2}$. In other words, D–policy is superior to N–policy if and only if $\rho \geq \frac{1}{2}$.

We now consider a different special case where the service times are deterministic, given by a constant C. In this case, the comparison between the three control policies is fairly easy. From the definition of the D–policy, the server is turned on whenever the workload in the system exceeds a predetermined level D. Since the service times are deterministic, this is equivalent to saying that the server is turned on as soon as the $(|C/D|+1)^{th}$ customer in a busy period arrives, where the operation $|t|$ is defined as taking the integer part of t. Because $(|C/D|+1)$ is an integer, a system operating under the D–policy is identical to one operating under the N–policy. In fact, this conclusion is valid under any cost structure and for any arrival process (see Balachandran and Tijms [1975], where they claimed incorrectly that the D–policy is superior to the N–policy for deterministic services).

Discussion From the above analysis, we conclude that the D–policy is not always better than the N–policy, a conclusion different from that of Balachandran and Tijms [1975] and Boxma [1976]. This is due to the difference between the cost structure we use and the one by Balachandran [1973], in the nature of the third cost component described at the beginning of this section. We consider this cost component as a linear function of

customer delay, whereas Balachandran [1973] a linear function of time–average workload. Under the condition that the server is always on (as in the standard M/G/1 queue case), the time–average workload is equivalent to customer delay, because of the PASTA (Poisson Arrivals See Time Average) property of Poisson arrivals (Wolff [1982]); and there is no difference between these two cost assumptions although they are defined differently. When server can be turned off, as required by the operation rules of all control policies, the time–average workload in the system is no longer equal to the customer delay. In fact, the expected customer delay is larger than the mean time–average workload in the system; the difference is $\frac{1}{\lambda E(|K|)}\sum_{n=1}^{\infty} nG^{[n]}(D)$ in the D–policy case (compare (2.23) with the expected time–average workload given in Balachandran and Tijms [1975]), and is $(1-\rho)\frac{N-1}{2\lambda}$ in the N–policy case. These terms are due to the average delay to those who arrived during off–periods, caused by server absence. Since without this term, the customer delay would be equivalent to the time–average workload, it is easy to see that the relative "weight" of this term in the cost function determines the relative performance of the control policy being used. We note that in the N–policy case, this delay depends only on the interarrival times while in the D–policy case, it depends not only on the interarrival times but also on the service times. Intuitively, when the influence of service times is negligible, the relative "weight" of this "extra" delay would also be negligible, a case close to the one studied by Balachandran. In the exponential service case, as we have seen, when traffic is heavy, the service times do not play an significant role; and it turns out that the D–policy is superior to the N–policy. When traffic is light, the conclusion is just the opposite. The same reasoning suggests that for deterministic services, there should be no difference in long–run average cost between the D–policy and the N–policy because the "randomness" of the service times vanishes. This reasoning also turns out to be true. We conjecture that this argument holds for all service distributions, that is, in the cost structure proposed by Yadin and Naor [1963], the D–policy tends to outperform the N–policy when traffic is heavy, and the reverse holds when traffic is light.

CHAPTER 3

COMPUTATIONAL ANALYSIS OF THE C/G/1/K QUEUE

3.1 INTRODUCTION

As mentioned in Chapter one, almost all queues in reality have limited capacities. Only for queues whose capacities are so large that the fraction of lost customers is negligible can the infinite–capacity approximation be justified. Consequently, many research results have been developed in the literature for the basic and important M/G/1/K model (see, for example, Cohen [1982], Chapter III; and Keilson [1966]). Since the joint probability distribution of queue length and remaining duration of the service time, if any, in progress at customer arrival epochs, is the basic information needed for derivation of other quantities of interest in this model, it is typical to define the queue length and remaining service time as "state variables" and then study the resulting Markov process. Such analyses were complicated by the fact that a continuous state space is needed to monitor the remaining duration of the service time in progress. Recently, Niu and Cooper [1989] reexamined the M/G/1/K model and proposed a new method of analysis which differs from the traditional approach in that they work with a more detailed (continuous) state space that is judiciously designed to include enough information so that the solution of the problem could eventually be reduced to that of a discrete–state Markov chain embedded immediately after service starts, monitoring the number of customers in queue. They also demonstrated that it is possible to solve the finite–capacity and infinite–capacity M/G/1 queues by a unified method. Moreover, since their method is based on "sample averages", all of their results have explicit term–by–term probabilistic interpretations.

Another important advantage of their method is that the analysis does not critically depend on the assumption of Poisson arrivals. In fact, it is not even necessary for the arrival and service processes to be of the usual renewal type. *All that is needed is "sufficient exponentiality" in the arrival process so that a suitably designed embedded discrete–state Markov chain exists immediately after service–start epochs.* This statement implies that the method introduced by Niu and Cooper [1989] is well suited for incorporating continuous–time–Markov–chain governed arrival processes. More specifically, it is easy to see that a natural way to introduce dependence in the arrival process is to augment the state space of the service–start embedded Markov chain to include one additional parameter to monitor the "status" of the arrival process. It is also natural to assume that the probabilistic behavior of future arrivals from any given time t onwards can be determined completely if the current "status" of the arrival process, which we take to be the state of the environmental Markov chain, is given. We will call arrival processes satisfying these conditions C–processes.

Obviously, C–processes offer great flexibility in terms of characterizing and/or approximating general arrival processes in practical applications. Because of its generality, it is not clear how traditional methods might apply to the analysis of the C/G/1/K model. The purpose of our work in this chapter is to analyze the probabilistic behavior of the C/G/1/K queue. Because our method of analysis is similar to the one used by Niu and Cooper [1989], the main focus here is on computational aspects of this model.

The outline of this chapter is as follows. In Section 3.2, we discuss our basic sample–average method and apply it to solve the C/G/1/K model in its most general form. In particular, we derive the long–run average behavior of the number of customers in the system and the remaining service time, if any, in progress, as observed by arriving customers. In Section 3.3, we give a detailed discussion of computational issues. Complete

procedures for the computation of the joint distribution of queue length and the remaining service times is given in this section. We also discuss how necessary computations can be simplified when the service time distribution is either generalized hyperexponential or Erlang.

3.2 THE C/G/1/K MODEL AND ITS ANALYSIS

3.2.1 The C/G/1/K Model

Denote by $\mathbf{X} \equiv \{X(t), t \geq 0\}$ the continuous–time environmental Markov chain, with a finite state space $\mathbb{N}$. The process $\mathbf{X}$ has the properties that each time it enters state i ($i \in \mathbb{N}$): (1) the amount of time it spends in that state before making the next transition is exponentially distributed with rate ν_i; and (2) when the process leaves state i, it will next enter state j with probability p_{ij}, where $\sum_{j \in \mathbb{N}} p_{ij} = 1$. An arrival process is called a C–process if the probabilistic behavior of future arrivals can be determined completely from the current state of the environmental Markov chain $\mathbf{X}$.

In the C/G/1/K model, the system has a total capacity of K customers including the one (if any) in service. An arriving customer enters the system only if it is not at full capacity; otherwise, he is lost immediately without receiving any service. Entering customers are served according to the first–in–first–out service discipline, without preemptions.

We will say that the state of the system is (j,x) if the number of customers in the system equals j ($j \geq 1$) and the remaining duration of the service in progress is less than or equal to x ($x \geq 0$). For $1 \leq j \leq K$ and $x \geq 0$, denote by $\alpha^j(x)$ the limiting proportion of customers who, on their arrival, find the system in state (j,x). We also define α^0 as the limiting proportion of customers who, on their arrival, find the system empty. α^0 and

$\alpha^j(x)$ together describe the limiting joint state behavior as observed by arriving customers. Given α^0 and $\alpha^j(x)$, we can easily derive other information about the behavior of the system. For example, the distribution of customer delay can be determined from α^0 and $\alpha^j(x)$:

$$P(W \le x) = \frac{1}{1 - \alpha^K(\infty)} \left[\alpha^0 + \sum_{j=1}^{K-1} \int_0^x G^{[j-1]}(x-t)\, d\alpha^j(x)\right], \qquad (3.1)$$

where $[1 - \alpha^K(\infty)]$ is the probability that a randomly selected customer actually enters the system. (3.1) is derived by conditioning on the state of the system at the arrival epoch of a randomly selected entering customer. Thus, the analysis of α^0 and $\alpha^j(x)$ plays an important role in the study of C/G/1/K queues.

3.2.2 The Service–Start Markov Chain And the Arrival Pattern in Service Intervals

We first consider the C/G/1/K queue. Suppose a customer is randomly selected from an infinite poll of customers. At the arrival epoch of this customer, he will find the system to be either empty or containing j ($1 \le j \le K$) customers. In the latter case, the waiting customers arrived either before the start or during the "age" of the current service. Define the customers in queue who arrived before the start of the current service as type–one customers and those who arrived during the "age" of the current service type–two, *including those who did not enter the system.* It is easy to see that the number of type–one customers in the system immediately after service–start epochs is governed by a Markov chain. The state space of this Markov chain is $\{(m,i),\ m \in \mathbb{N},\ i = 0', 0, 1, 2,..., K-2\}$, where the first component m is interpreted as the state of the environmental Markov chain of the arrival process, and the second component the number of customers in queue, with the additional stipulation that the service is the first one in a busy period if and only if i=0'. Because the number of type–two customers in the system is independent of the number of type–one customers, we can then decompose our analysis of α^0 and $\alpha^j(x)$ into

three basic parts: (1) determine α^0, the fraction of customers who find the system empty; (2) determine the distribution of type–one customers by solving a service–start Markov chain; and (3) determine the probability for a randomly selected arrival to find (i) j type–two customers prior to the arrival and (ii) the remaining duration of the service time in progress less than or equal to x. Because the evaluation of α^0 requires the stationary distribution of the service–start Markov chain, we will analyze this Markov chain first. Before continuing, we introduce the following notation:

$$N_i(t) : \text{the number of arrivals in a time interval } [0, t] \text{ given that } X(0) = i, \quad (3.2a)$$

$$T_i^n : \text{time until the } n^{th} \text{ arrival given that } X(0) = i, \quad (3.2b)$$

$$P_{ij}(t) = P[X(t) = j \mid X(0) = i], \quad (3.2c)$$

$$P_{ij}(n,t) = P[X(t) = j, N_i(t) = n \mid X(0) = i], \quad (3.2d)$$

$$A_i^n(t) = P[T_i^n \le t], \quad (3.2e)$$

$$Q_{ij} = P[X(T_i^1) = j]. \quad (3.2f)$$

Thus, $P_{ij}(n,t)$ is the probability of the event that starting from state i, the environmental Markov chain will be in state j at time t, and there are n arrivals during this time period (it is easy to see that $P_{ij}(t) = \sum_{n=0}^{\infty} P_{ij}(n,t)$); $A_i(n,t)$ is the probability that given the environmental Markov chain is in state i at time 0, the n^{th} customer will arrive before time t; and Q_{ij} is the probability of the event that starting from state i, the environmental Markov chain will be in state j when the next customer arrives. $P_{ij}(n,t)$, $A_i(n,t)$, and Q_{ij} are basic quantities associated with the arrival process. We will treat these probabilities as input information to keep the analysis of α^0 and $\alpha^j(x)$ as concise as possible. Computation of these probabilities will be discussed independently in the next section.

Denote by σ_m^i the stationary state distribution of the service–start embedded Markov chain, where the superscript stands for the number of customers present in the queue immediately after a service start and the subscript m for the state of the

environmental Markov chain. For $m \in \mathbb{N}$ and $i = 0', 0, 1, 2, \ldots, K-2$, σ_m^i is determined by

$$\sigma_m^{0'} = \sum_{k\in\mathbb{N}} \sum_{n\in\mathbb{N}} (\sigma_k^{0'} + \sigma_k^0)\, a_{kn}^0\, Q_{nm}\,, \tag{3.3a}$$

$$\sigma_m^j = \sum_{k\in\mathbb{N}} \sum_{i=0'}^{j+1} \sigma_k^i\, a_{km}^{j+1-i} \qquad \text{for } 0 \le j < K-3\,, \tag{3.3b}$$

$$\sigma_m^{K-2} = \sum_{k\in\mathbb{N}} \sum_{i=0'}^{K-2} \sigma_k^i \sum_{j=K-2}^{\infty} a_{km}^{j+1-i}, \tag{3.3c}$$

$$\sum_{k\in\mathbb{N}} \sum_{i=0'}^{K-2} \sigma_k^i = 1 \tag{3.3d}$$

where a_{km}^j is an entry of the transition probability matrix of the service–start Markov chain, and is given by

$$a_{km}^j = \int_0^\infty P_{km}(j,t)\, dG(t). \tag{3.4}$$

(3.3) is obtained from standard stationary equations of the form $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$. More specifically, it is derived based on the observation that the number of customers left behind by a service start is one less than the sum of (i) the number of customers left behind by the previous service start and (ii) those who arrived during this time period (except when the service start initiates a busy period). The solution of (3.3) is unique and can be solved by standard methods. In our later analysis, we will simply treat σ_m^j as given.

We next evaluate the conditional joint probability of an arriving customer finding j type–two customers in the system and the remaining duration of the service in progress to be less than or equal to x, given that he arrives in a "type–m service period" (defined to be a service period that starts with the environmental Markov chain in state m). We will denote this joint probability by $\nu_m^j(x)$. To evaluate $\nu_m^j(x)$, we consider a discrete renewal reward process whose "interevent times" are defined to be the numbers of arrivals in type–m service periods, and which earns a reward of one unit whenever an arrival finds j type–two customers in the queue and the remaining duration of the service in progress is

less than or equal to x. Then, from standard renewal reward theory, $\nu_m^j(x)$ is given as the ratio of the expected reward in a typical renewal cycle to its expected "length". That is,

$$\nu_m^j(x) = \frac{1}{E(N_m)} \int_0^\infty [G(t+x) - G(t)]\, dA_m^{j+1}(t), \tag{3.5}$$

where N_m denotes the number of arrivals in a type–m service period. The integral in (3.5) (giving the expected reward) is obtained by conditioning on the arrival time of the $(j+1)^{th}$ customer after the start of a type–m service period (see Niu and Cooper [1989] for related discussions).

Treating σ_m^j and $\nu_m^j(x)$ as given, we are now ready to derive α^0 and $\alpha^j(x)$ for our C/G/1/K system, which we do in the next section.

3.2.3 Derivation of α^0 And $\alpha^j(x)$ by "Successive Relative Sampling"

The derivations of σ_m^j and $\nu_m^j(x)$ in the above subsection were based on the information on the "status" of the environmental Markov chain at the beginning of each service period. To simplify our analysis, it turns out to be easier to work with $\alpha_m(i,j,x)$, $0 \le i \le K-1$, $j \ge 0$, and $x \ge 0$, defined as the limiting proportion of customers who, on their arrival, find that: (1) there are i type–one and j type–two customers in the system respectively; (2) the remaining duration of the service in progress is less than or equal to x; and (3) the environmental Markov chain was in state m at the start of the current service. We also define α_m^0 in a similar way. α_m^0 and $\alpha_m(i,j,x)$ are related to α^0 and $\alpha^j(x)$ by

$$\alpha^0 = \sum_{m \in \mathbb{N}} \alpha_m^0 , \tag{3.6a}$$

$$\alpha^j(x) = \sum_{m \in \mathbb{N}} \left[\sum_{i=0}^{j-1} \alpha_m(i,j-1-i,x) \right], \tag{3.6b}$$

and

$$\alpha^K(x) = \sum_{m \in \mathbb{N}} \sum_{j=K}^{\infty} \left[\sum_{i=0}^{j-1} \alpha_m(i,j-1-i,x) \right]. \tag{3.6c}$$

The type–two customers counted by $\alpha_m(i,j-1-i,x)$ for $j \geq K$ in the right–hand side of (3.6c) include those who are lost after the start of the current service–time.

Observe that the proportion of arriving customers that are lost without receiving any service is given by $\alpha^K(\infty)$; and, on the other hand, the proportion of those who actually enter the system is given by $1 - \alpha^K(\infty)$. Since every *entering customer* who finds the system empty and the environmental Markov chain in state m *triggers an (m,0') type service*, we therefore have

$$\frac{\alpha_m^0}{1 - \alpha^K(\infty)} = \sigma_m^{0'},$$

leading to

$$\alpha^0 = \sum_{m \in \mathbb{N}} \alpha_m^0 = [1 - \alpha^K(\infty)] \sum_{m \in \mathbb{N}} \sigma_m^{0'}. \tag{3.7}$$

Our derivation of $\alpha_m(i,j,x)$ will be based on a successive relative sampling scheme. Let the original sample space be the set of all arriving customers; then the steps are as follows:

Step 1. Select a customer randomly from the original sample space. The proportion of customers in the sample space that are blocked on their arrival is given by $1 - \alpha^0$.

Step 2. Select a customer randomly from the set of *blocked customers.* The relative proportion of customers from this subset of the original sample space that will "interrupt" a "type–m" service period is given by

$$\frac{E(N_m)\Sigma_i \sigma_m^i}{\Sigma_m [E(N_m)\Sigma_i \sigma_m^i]}.$$

This expression is derived by considering explicitly biases caused by different initial conditions at the beginning of different types of service periods. More specifically, we view the number of arrivals in successive service period as "sojourns" in a discrete "time"

semi–Markov process and apply Theorem 4.8.3 of Ross [1983] to derive this expression. To simplify later expressions, let E(N) be $\Sigma_m[E(N_m)\Sigma_i\sigma_m^i]$.

Step 3. Select a blocked customer randomly from those who "interrupt" a type–m service period. The relative proportion of customers in this further divided subset of customers that will find i type–one customers in the system on their arrival is given by

$$\frac{E(N_m)\sigma_m^i}{\Sigma_i E(N_m)\sigma_m^i} = \frac{\sigma_m^i}{\Sigma_i \sigma_m^i}$$

and that of those who find j type–two customers in the system and the remaining duration of the service in progress less than or equal to x is given by $\nu_m^j(x)$.

Since the sample space in each step of the above sampling scheme is a subspace of the one in the previous step, it is clear that the proportion of customers in the original sample space that: (1) are blocked on their arrival (step 1); (2) interrupt a type–m service period (step 2); (3) find i type–one customers in the system; and (4) find j type–two customers in the system and the remaining duration of the service in progress less than or equal to x (step 3) is given by

$$(1-\alpha^0) \times \frac{E(N_m)\Sigma_i\sigma_m^i}{E(N)} \times \frac{\sigma_m^i}{\Sigma_i \sigma_m^i} \times \nu_m^j(x). \tag{3.8}$$

The last two terms in the above expression are related in product form because (as mentioned earlier) the number of type–two customers in the system as seen by a randomly selected arrival is independent of that of the type–one customers. This expression gives $\alpha_m(i,j,m)$ for $1 \le i+j \le K-1$. Rewrite $\nu_m^j(x)$ in (3.5) as $\theta_m^j(x)/E(N_m)$, where

$$\theta_m^j(x) = \int_0^\infty [G(t+x) - G(t)]\, dA_m^{j+1}(t). \tag{3.9}$$

Substituting this expression into (3.6b) and (3.6c), we get, for $1 \le j \le K-1$,

$$\alpha^j(x) = \frac{1}{E(N)}(1-\alpha^0) \sum_{m\in\mathbb{N}} \left[\sigma_m^{0'}\theta_m^{j-1}(x) + \sum_{k=0}^{j} \sigma_m^k \theta_m^{j-k-1}(x)\right],$$

and

$$\alpha^K(x) = \frac{1}{E(N)}(1-\alpha^0) \sum_{m\in\mathbb{N}} \sum_{j=K}^{\infty} \left[\sigma_m^{0'}\theta_m^{j-1}(x) + \sum_{k=0}^{K-2} \sigma_m^k \theta_m^{j-k-1}(x)\right].$$

These two expressions together with (3.7) enable us to write our final results as

$$\alpha^0 = \sum_{m\in\mathbb{N}} \sigma_m^{0'}[1-\alpha^K(\infty)], \tag{3.10a}$$

$$\alpha^j(x) = \frac{1}{E(N)} \sum_{m\in\mathbb{N}} \left[1 - \sum_{m\in\mathbb{N}} \sigma_m^{0'}[1-\alpha^K(\infty)]\right] \left[\sigma_m^{0'}\theta_m^{j-1}(x) + \sum_{k=0}^{j} \sigma_m^k \theta_m^{j-k-1}(x)\right], \tag{3.10b}$$

for $1 \le j \le K-1$, and

$$\alpha^K(x) = \frac{1}{E(N)} \sum_{m\in\mathbb{N}} \left[1 - \sum_{m\in\mathbb{N}} \sigma_m^{0'}[1-\alpha^K(\infty)]\right] \times \left[\sigma_m^{0'} \sum_{j=K}^{\infty} \theta_m^{j-1}(x) + \sum_{k=0}^{K-2} \sum_{j=K}^{\infty} \sigma_m^k \theta_m^{j-k-1}(x)\right]. \tag{3.10c}$$

The only unknown $\alpha^K(\infty)$ in (3.10) can be determined by solving (3.10c) after passing x to ∞, leading to

$$\alpha^K(\infty) = \frac{(1 - \Sigma_m \sigma_m^{0'})\Sigma_m \beta_m}{E(N) - (\Sigma_m \sigma_m^{0'})\Sigma_m \beta_m}, \tag{3.11a}$$

where

$$\beta_m = \sum_{j=K}^{\infty} \left[\sigma_m^{0'}\theta_m^{j-1}(\infty) + \sum_{k=0}^{K-2} \sigma_m^k \theta_m^{j-k-1}(\infty)\right]. \tag{3.11b}$$

3.3 COMPUTATION OF α^0 AND $\alpha^j(x)$

In the previous section, we described our basic method for deriving α^0 and $\alpha^j(x)$ for the C/G/1/K queue. No specific structure of the arrival process other than the environmental Markov chain is required. Although many point processes can be viewed as C–processes, the difference between these processes are reflected only through the three basic probabilities $P_{ij}(n,t)$, $A_i^n(t)$ and Q_{ij}, which were taken as input information in the previous section. In this section, we will first consider the case of the doubly stochastic Poisson input as an example, to illustrate the general approach of evaluating these probabilities (Section 3.3.1). We then discuss computational procedures for α^0 and $\alpha^j(x)$ in general C/G/1/K systems (Section 3.3.2). A numerical example is given in Section 3.3.3. Finally, in Section 3.3.4, we discuss more efficient procedures for solving α^0 and $\alpha^j(x)$ when the service–time distribution is either generalized hyperexponential or Erlang.

3.3.1 Evaluation of $P_{ij}(n,t)$, $A_i^n(t)$, and Q_{ij}

From the definitions of $P_{ij}(n,t)$, $A_i^n(t)$, and Q_{ij}, it is clear that these quantities are independent of any specific queueing models. Solving for these probabilities is a standard problem.

We first derive the derivative of $A_i^n(t)$ (because only the derivative of $A_i^n(t)$ is needed in our analysis, see expressions (3.10) and (3.9)). To do this, we note that if the n^{th} customer arrives in the interval (t,t+dt), then there must be n–1 arrivals in the time interval (0,t). Hence, by conditioning on the state of the environmental Markov chain at time t, we obtain

$$dA_i^n(t) = \Sigma_m P_{im}(n-1,t)\lambda_m dt. \tag{3.12}$$

Therefore, relation (3.12) reduces the problem of evaluating $dA_i^n(t)$ to that of evaluating

$P_{ij}(n,t)$.

To compute the probabilities $P_{ij}(n,t)$ and Q_{ij} for i, j $\in \mathbb{N}$, we specialize our C–process to a doubly stochastic Poisson arrival process to illustrate the general approach of computing these probabilities.

In a doubly stochastic Poisson process, customers arrive according to a Poisson process whose rate is governed by an continuous–time Markov chain. When the continuous–time Markov chain is in state i (i$\in\mathbb{N}$), we let the arrival rate be λ_i.

To compute Q_{ij} for this arrival process, we condition on whether a customer arrives before of after the environmental Markov chain changes its state. This leads to a system of linear equations

$$Q_{ij} = \frac{\nu_i}{\lambda_i + \nu_i} \Sigma_m \, p_{im} Q_{mj} + \delta_{ij} \frac{\lambda_i}{\lambda_i + \nu_i}, \qquad \text{for } i, j \in \mathbb{N}, \tag{3.13}$$

where δ_{ij} equals 1 or 0 depending on whether i=j or not. Since the subscripts of Q_{ij} may take any value from the index set $\mathbb{N}$, this system of linear equations is of order $|\mathbb{N}|^2$.

Next, we focus on the computation of the probability $P_{ij}(n,t)$. We first state the boundary conditions for $P_{ij}(n,t)$, i, j $\in \mathbb{N}$, as

$$\lim_{t \to 0} P_{ij}(n,t) = \delta_{n0}\delta_{ij}. \tag{3.14}$$

By conditioning on the time at which the environmental Markov chain changes its state, we have, for i, j $\in \mathbb{N}$ and n = 0,1,...,K–1,

$$P_{ij}(n,t) = \int_0^t \sum_{k=0}^{n} \left[\frac{(\lambda_i u)^k}{k!} e^{-\lambda_i u} \Sigma_m p_{im} P_{mj}(n-k,t-u)\right] e^{-\nu_i u} \nu_i du + \delta_{ij}\left[\frac{(\lambda_i t)^n}{n!} e^{-\lambda_i t}\right] e^{-\nu_i t} \tag{3.15}$$

The probabilistic interpretation of this expression is as follows: if k customers arrive before the first transition, at time u, of the environmental Markov chain, then the remaining n–k

customers must arrive in the time interval (t–u,t).

The reason for setting the range of n in $P_{ij}(n,t)$ to $0 \le n \le K-1$ is that we want to control the computational effort needed for solving this finite–capacity queue as the capacity of the queue increases. It is desirable to carry out our analysis with quantities that count arrivals only up to K. In fact, if this is done throughout, the computational effort involved in solving finite–capacity queues is less than that of solving corresponding larger capacity queues.

Differentiating (1,15) with respect to t, we get

$$\frac{d}{dt}P_{ij}(n,t) = \sum_{k=0}^{n}\left[\frac{(\lambda_i u)^k}{k!}e^{-\lambda_i u}\Sigma_m p_{im}\, P_{mj}(n-k,t-u)\nu_i e^{-\nu_i u}\right]_{u=t} +$$

$$+ \int_0^t \sum_{k=0}^{n}\left[\frac{(\lambda_i u)^k}{k!}e^{-\lambda_i u}\Sigma_m p_{im}\,\frac{d}{dt}\left[P_{mj}(n-k,t-u)\right]\right]\nu_i e^{-\nu_i u}du -$$

$$-(\lambda_i+\nu_i)\delta_{ij}\left[\frac{(\lambda_i t)^n}{n!}e^{-\lambda_i t}\right]e^{-\nu_i t} + \lambda_i\delta_{ij}\left[\frac{(\lambda_i t)^{n-1}}{(n-1)!}e^{-\lambda_i t}\right]e^{-\nu_i t}.$$

Substituting the relation $\frac{d}{dt}P_{mj}(n,t-u) = -\frac{d}{du}P_{mj}(n,t-u)$ into the above equation and integrating the resulting system by parts, we obtain

$$\frac{d}{dt}P_{ij}(n,t) = \sum_{k=0}^{n}\left[\frac{(\lambda_i u)^k}{k!}e^{-\lambda_i u}\Sigma_m p_{im}\, P_{mj}(n-k,t-u)\nu_i e^{-\nu_i u}\right]_{u=t} -$$

$$-\sum_{k=0}^{n}\left[\frac{(\lambda_i u)^k}{k!}e^{-\lambda_i u}\,\Sigma_m p_{im}P_{mj}(n-k,t-u)\nu_i e^{-\nu_i u}\right]_{u=0}^{u=t} +$$

$$+\int_0^t \nu_i \sum_{k=0}^{n}\Sigma_m p_{im}P_{mj}(n-k,t-u)\, d\left[\frac{(\lambda_i u)^k}{k!}e^{-(\lambda_i+\nu_i)u}\right] -$$

$$-(\lambda_i+\nu_i)\delta_{ij}\left[\frac{(\lambda_i t)^n}{n!}e^{-\lambda_i t}\right]e^{-\nu_i t} + \lambda_i\delta_{ij}\left[\frac{(\lambda_i t)^{n-1}}{(n-1)!}e^{-\lambda_i t}\right]e^{-\nu_i t}$$

$$= \nu_i\Sigma_m p_{im}P_{mj}(n,t) -$$

$$- (\lambda_i+\nu_i)\int_0^t \sum_{k=0}^{n}\left[\frac{(\lambda_i u)^k}{k!}e^{-\lambda_i u}\Sigma_m p_{im}P_{mj}(n-k,t-u)\right]e^{-\nu_i u}\nu_i\, du +$$

$$+ \lambda_i\int_0^t \sum_{k=1}^{n}\left[\frac{(\lambda_i u)^{k-1}}{(k-1)!}e^{-\lambda_i u}\Sigma_m p_{im}P_{mj}(n-k,t-u)\right]e^{-\nu_i u}\nu_i\, du -$$

$$- (\lambda_i+\nu_i)\delta_{ij}\left[\frac{(\lambda_i t)^n}{n!}e^{-\lambda_i t}\right]e^{-\nu_i t} + \lambda_i\delta_{ij}\left[\frac{(\lambda_i t)^{n-1}}{(n-1)!}e^{-\lambda_i t}\right]e^{-\nu_i t}.$$

By grouping terms and applying (3.15), we obtain, for i, j $\in \mathbb{N}$ and $0 \le n \le K-1$,

$$\frac{d}{dt}P_{ij}(n,t) = \nu_i\Sigma_m p_{im}P_{mj}(n,t) - (\lambda_i+\nu_i)P_{ij}(n,t) + \lambda_i P_{ij}(n-1,t), \tag{3.16}$$

where we have assumed that $P_{ij}(n-1,t) = 0$ when $n = 0$.

(3.13) and (3.16) are derived for a doubly stochastic Poisson process. But the method of analysis is valid for all C–processes.

(3.15) and (3.16) are actually systems of linear integral equations and linear differential equations respectively. Since the unknowns in (3.15) are conditional probabilities, the system must have a unique solution. On the other hand, (3.16) is derived from (3.15). Its solution must be the same as that of (3.15). In other words, if one of the systems has a solution, it must also be the solution of the other system. We write (3.16) in the matrix form

$$\frac{d}{dt}\mathbf{X}_j = \mathbf{A}\mathbf{X}_j,$$

where $\mathbf{X}_j = (P_{1j}(0,t), P_{2j}(0,t),..., P_{1j}(1,t), P_{2j}(1,t),...)^T$, $j \in \mathbb{N}$; $\mathbf{A}$ is a $(K\times|\mathbb{N}|)$ by $(K\times|\mathbb{N}|)$ matrix. From the theory of differential equations, *the solution* of *this system is completely and uniquely determined by the eigenvalues of the coefficient matrix A and the boundary conditions* (3.14). Since the system $\frac{d}{dt}\mathbf{X}_j = \mathbf{A}\mathbf{X}_j$ is identical for every $j\in \mathbb{N}$, the general solution (i.e. solutions without using boundary conditions) for different $\mathbf{X}_j$, $j\in \mathbb{N}$ will be identical. Therefore, to solve (3.16), we only need to solve one of the systems:

$\frac{d}{dt}\mathbf{X}_j = \mathbf{A}\mathbf{X}_j$, $j \in \mathbb{N}$ and then use boundary conditions to determine $\mathbf{X}_j$ for every $j \in \mathbb{N}$.

Computationally, solving a system of form $\frac{d}{dt}\mathbf{X} = \mathbf{A}\mathbf{X}$ with given boundary conditions involves calculation of γ_i and $\mathbf{v}_i$, the eigenvalues and eigenvectors of matrix $\mathbf{A}$ for $i = 1, 2, \ldots, K\times|\mathbb{N}|$. First, we consider the case that all the eigenvalues are real and distinct. (Complex eigenvalues lead to trigonometric–function solution, which cannot occur in systems whose solutions are probability distributions because of the nonnegativity condition). According to differential equation theory, the solution of the system $\frac{d}{dt}\mathbf{X} = \mathbf{A}\mathbf{X}$ is given by

$$\mathbf{X}(t) = \Sigma_k \mathbf{v}_k e^{\gamma_k t}. \tag{3.17}$$

As a quick check, we can substitute (3.17) into the system $\frac{d}{dt}\mathbf{X} = \mathbf{A}\mathbf{X}$, leading to

$$\frac{d}{dt}\mathbf{X} = \Sigma_k \gamma_k \mathbf{v}_k e^{\gamma_k t} = \Sigma_k \mathbf{A}\mathbf{v}_k e^{\gamma_k t},$$

or equivalently, $\Sigma_k e^{\gamma_k t}[\mathbf{A} - \gamma_k \mathbf{I}]\mathbf{v}_k = \mathbf{0}$, which is exactly the weighted sum of the charac–teristic equations of matrix $\mathbf{A}$. Thus, (3.17) satisfies the differential equation system and is indeed the solution. The Euclidean norms of the eigenvectors in (3.17) are so far arbitrary. They serve as the free constants required in the general solution of the differential equation system $\frac{d}{dt}\mathbf{X} = \mathbf{A}\mathbf{X}$. These free constants will be determined by boundary conditions.

We now consider the case that not all eigenvalues are distinct; and suppose γ is an n–fold eigenvalue of matrix $\mathbf{A}$, and $\mathbf{v}$ is the corresponding eigenvector of γ. In this case, the solution determined by γ will no longer be $\mathbf{v}\cdot e^{\gamma t}$, instead, it takes the form

$$(\mathbf{b}_0 + \mathbf{b}_1 t + \mathbf{b}_2 t^2 + \cdots \cdots + \mathbf{b}_{n-1} t^{n-1}) e^{\gamma t}. \tag{3.18}$$

Substituting (3.18) into the system $\frac{d}{dt}\mathbf{X} = \mathbf{A}\mathbf{X}$, we obtain

$$\gamma(b_0 + b_1 t + b_2 t^2 + \cdots + b_{n-1} t^{n-1}) e^{\gamma t} + (b_1 + 2b_2 t + \cdots + (n-1) b_{n-1} t^{n-2}) e^{\gamma t}$$

$$= A(b_0 + b_1 t + b_2 t^2 + \cdots \cdots + b_{n-1} t^{n-1}) e^{\gamma t}.$$

Since this equation holds for all t, we have

$$\begin{aligned}
(A - \gamma I) b_0 &= b_1, \\
(A - \gamma I) b_1 &= 2b_2, \\
(A - \gamma I) b_2 &= 3b_3, \\
&\cdots \cdots \\
(A - \gamma I) b_{n-2} &= (n-1) b_{n-1}, \\
(A - \gamma I) b_{n-1} &= 0 .
\end{aligned} \tag{3.19}$$

We see, from (3.19), that $\mathbf{v}$, the eigenvector corresponding to γ, is the solution of b_{n-1}; and all other coefficient vectors are recursively determined by $\mathbf{v}$. Again, the Euclidean norm of $\mathbf{v}$ is determined by boundary conditions. For a complete discussion of solution methods for systems of linear differential equations, please refer to Chapter 7 of Brauer, Nohel, Schneider [1976].

3.3.2 Computational Procedure for α^0 and $\alpha^j(x)$

So far, we have evaluated all the quantities required for the computation of α^0 and $\alpha^j(x)$ for the C/G/1/K queue for j = 1, 2, ..., K. A step–by–step computational procedure is now summarized as follows:

Step 1: Solve (3.16) for $P_{ij}(n,t)$, $i, j \in \mathbb{N}$ and n = 0, 1, ..., K–1. In this step, we need to compute all eigenvalues and eigenvectors of the coefficient matrix A of the linear differential equation system $\frac{d}{dt}X_j = AX_j$, which is of order $K \times |\mathbb{N}|$. We also need to solve a system of linear equations to determine the Euclidean norms of all eigenvectors.

Step 2: Solve (3.13) for Q_{ij}, i, j $\in \mathbb{N}$ and then compute a^j_{km} for $0 \le j \le K-1$ and m, k $\in \mathbb{N}$ using (3.4). The evaluation of the integral in (3.4) constitutes the main computational effort needed in this step.

Step 3: Solve (3.3) for σ^j_m, m $\in \mathbb{N}$ and j = 0', 0, 1,..., K–2.

Step 4: Compute $\theta^j_m(x)$ for m $\in \mathbb{N}$ and $0 \le j \le K-1$ from relations (3.12) and (3.9), and determine α^0 and $\alpha^j(x)$ from (3.10)

Step 5: Compute α^K from (3.11), while treating E(N) as a known constant.

Step 6: Determine E(N) by the normalization condition $\alpha^0 + \Sigma_j \alpha^j(\infty) = 1$.

Note that in Step 3, if we try to compute σ^i_k by (3.3), we need to compute a^j_{km} for j > K–1, which in turn requires knowledge of $P_{km}(j,t)$ for j > K–1. As mentioned earlier in the previous subsection, we do not calculate $P_{km}(j,t)$ for j > K–1 in our procedure. Therefore, we must express (3.3c) in terms of information that is already available. Here, we use (3.4) to express (3.3c) as

$$\begin{aligned}\sigma^{K-2}_m &= \sum_{k\in\mathbb{N}} \sum_{i=0'}^{K-2} \sigma^i_k \sum_{j=K-2}^{\infty} a^{j+1-i}_{km} \\ &= \sum_{k\in\mathbb{N}} \sum_{i=0'}^{K-2} \sigma^i_k \sum_{j=K-2}^{\infty} \int_0^{\infty} P_{km}(j+1-i,t)\, dG(t).\end{aligned}$$

By exchanging the order of summation and integration in the above equality and using the relation $P_{km}(t) = \sum_{j=0}^{\infty} P_{km}(j,t)$, we finally express (3.3c) as

$$\begin{aligned}\sigma^{K-2}_m &= \sum_{k\in\mathbb{N}} \sum_{i=0'}^{K-2} \sigma^i_k \int_0^{\infty} \sum_{j=K-2}^{\infty} P_{km}(j+1-i,t) dG(t) \\ &= \sum_{k\in\mathbb{N}} \sum_{i=0'}^{K-2} \sigma^i_k \int_0^{\infty} \Big[P_{km}(t) - \sum_{j=0}^{K-1} P_{km}(j+1-i,t) \Big] dG(t)\end{aligned}$$

$$= \sum_{k\in\mathbb{N}} \sum_{i=0'}^{K-2} \sigma_k^i \left[a_{km} - \sum_{j=0}^{K-1} a_{km}^{j+1-i} \right], \tag{3.20}$$

where

$$a_{km} = \int_0^\infty P_{km}(t)\, dG(t). \tag{3.21}$$

The probability $P_{km}(t)$ (see (3.2c)) satisfies a system of linear (Kolmogorov backward) differential equations given by

$$\frac{d}{dt} P_{ij}(t) = \Sigma_m \nu_i p_{im} P_{mj}(t) - \nu_i P_{ij}(t) \qquad \text{for } i, j \in \mathbb{N} \tag{3.22}$$

(see, for example, Ross [1983] pp. 147–151). The solution of (3.20) can be obtained by the same approach used earlier for solving (3.16).

For the same reason, we must rewrite (3.10c) so that it involves only $\theta_m^j(x)$ for $j \leq K-2$. To do this, we define

$$\theta_m(x) = \int_0^\infty [G(t+x) - G(t)]\, \Sigma_r P_{mr}(t) \lambda_r dt. \tag{3.23}$$

Note that

$$\begin{aligned} \sum_{j=0}^{\infty} \theta_m^j(x) &= \sum_{j=0}^{\infty} \int_0^\infty [G(t+x) - G(t)]\, \Sigma_r P_{mr}(j,t) \lambda_r dt \\ &= \int_0^\infty [G(t+x) - G(t)]\, \Sigma_r \left[\sum_{j=0}^{\infty} P_{mr}(j,t) \right] \lambda_r dt \\ &= \int_0^\infty [G(t+x) - G(t)]\, \Sigma_r\, P_{mr}(t) \lambda_r dt \\ &= \theta_m(x). \end{aligned} \tag{3.24}$$

With (3.24), (3.10c) can be expressed as

$$\alpha^K(x) = \frac{1}{E(N)} \Sigma_m \left[1 - [1 - \alpha^K]\, \Sigma_m \sigma_m^{0'} \right] \left[\sigma_m^{0'} \sum_{j\geq K}^{\infty} \theta_m^{j-1}(x) + \sum_{k=0}^{K-2} \sum_{j\geq K}^{\infty} \sigma_m^k \theta_m^{j-k-1}(x) \right]$$

$$= \frac{1}{E(N)} \Sigma_m \Big[1 - [1-\alpha^K]\,\Sigma_m \sigma_m^{0'}\Big]\Big[\sigma_m^{0'}\Big\{\theta_m(x) - \sum_{j=0}^{K-1} \theta_m^{j-1}(x)\Big\} + \sum_{k=0}^{K-2} \sigma_m^k \Big\{\theta_m(x) - \sum_{j=0}^{K-1} \theta_m^{j-k-1}(x)\Big\}\Big]. \tag{3.25}$$

Correspondingly, the expression of (3.11b) changes to

$$\beta_m = \sigma_m^{0'}\Big[\theta_m(\infty) - \sum_{j=0}^{K-1} \theta_m^{j-1}(\infty)\Big] + \sum_{k=0}^{K-2} \sigma_m^k \Big[\theta_m(\infty) - \sum_{j=0}^{K-1} \theta_m^{j-k-1}(\infty)\Big]. \tag{3.26}$$

To compute E(N), we pass x in (3.10a), (3.10b), and (3.25) to infinity and rewrite these expressions as

$$\alpha^0 = \alpha^0, \tag{3.27a}$$

$$\alpha^j(\infty) = \frac{1}{E(N)}(1-\alpha^0) \sum_{m \in \mathbb{N}} \Big[\sigma_m^{0'} \theta_m^{j-1}(\infty) + \sum_{k=0}^{j} \sigma_m^k \theta_m^{j-k-1}(\infty)\Big], \tag{3.27b}$$

for $1 \le j \le K-1$, and

$$\alpha^K(\infty) = \frac{1}{E(N)}(1-\alpha^0) \sum_{m \in \mathbb{N}} \Big[\sigma_m^{0'}\Big\{\theta_m(\infty) - \sum_{j=0}^{K-1} \theta_m^{j-1}(\infty)\Big\} + \sum_{k=0}^{K-2} \sigma_m^k \Big\{\theta_m(\infty) - \sum_{j=0}^{K-1} \theta_m^{j-k-1}(\infty)\Big\}\Big]. \tag{3.27c}$$

Summation of α^j over all j in (3.27) produces

$$\alpha^0 + \frac{1}{E(N)}(1-\alpha^0) \sum_{m \in \mathbb{N}} \Big[\sigma_m^{0'} + \sum_{k=0}^{K-2} \sigma_m^k\Big] \theta_m(\infty) = 1. \tag{3.28}$$

(3.28) implies that

$$E(N) = \sum_{m \in \mathbb{N}} \Big[\sigma_m^{0'} + \sum_{k=0}^{K-2} \sigma_m^k\Big] \theta_m(\infty) \quad . \tag{3.29}$$

3.3.3 A Numerical Example

In this subsection, we work with a specific example to illustrate the step–by–step computational procedure given in the previous subsection. In order to simplify the calcultion, we choose the capacity of the system to be 3, i.e. K = 3. We also restrict the state space of the environmental Markov chain to 2, i.e. $\mathbb{N} = \{1, 2\}$. In our example, the service–time is uniformly distributed in the interval of (0,1). All other parameters of the system are given as:

$$\lambda_1=0.9,\ \lambda_2=1.9,\ \nu_1=0.1,\ \nu_2=0.2;$$

$$p_{11}=0,\ p_{12}=1,\ p_{21}=0.5,\ p_{22}=0.5\ .$$

As described in the previous subsection, we first solve (3.16) for $P_{ij}(n,t)$, i, j = 1, 2 and n = 0, 1, 2. In our example, the system can be expressed as:

$$\frac{d}{dt}\begin{bmatrix} P_{11}(0,t) \\ P_{21}(0,t) \\ P_{11}(1,t) \\ P_{21}(1,t) \\ P_{11}(2,t) \\ P_{21}(2,t) \end{bmatrix} = \begin{bmatrix} -1 & 0.1 & 0 & 0 & 0 & 0 \\ 0.1 & -2 & 0 & 0 & 0 & 0 \\ 0.9 & 0 & -1 & 0.1 & 0 & 0 \\ 0 & 1.9 & 0.1 & -2 & 0 & 0 \\ 0 & 0 & 0.9 & 0 & -1 & 0.1 \\ 0 & 0 & 0 & 1.9 & 0.1 & -2 \end{bmatrix} \times \begin{bmatrix} P_{11}(0,t) \\ P_{21}(0,t) \\ P_{11}(1,t) \\ P_{21}(1,t) \\ P_{11}(2,t) \\ P_{21}(2,t) \end{bmatrix}, \tag{3.30a}$$

and

$$\frac{d}{dt}\begin{bmatrix} P_{12}(0,t) \\ P_{22}(0,t) \\ P_{12}(1,t) \\ P_{22}(1,t) \\ P_{12}(2,t) \\ P_{22}(2,t) \end{bmatrix} = \begin{bmatrix} -1 & 0.1 & 0 & 0 & 0 & 0 \\ 0.1 & -2 & 0 & 0 & 0 & 0 \\ 0.9 & 0 & -1 & 0.1 & 0 & 0 \\ 0 & 1.9 & 0.1 & -2 & 0 & 0 \\ 0 & 0 & 0.9 & 0 & -1 & 0.1 \\ 0 & 0 & 0 & 1.9 & 0.1 & -2 \end{bmatrix} \times \begin{bmatrix} P_{12}(0,t) \\ P_{22}(0,t) \\ P_{12}(1,t) \\ P_{22}(1,t) \\ P_{12}(2,t) \\ P_{22}(2,t) \end{bmatrix}. \tag{3.30b}$$

The eigenvalues of the coefficient matrix in either of these two subsystems are

$$\gamma_1 = \gamma_2 = \gamma_3 = -2.009902,$$

$$\gamma_4 = \gamma_5 = \gamma_6 = -0.990098.$$

Therefore, the solution corresponding to each distinct eigenvalue takes the form given by

(3.19). Following the procedure described in Section 3.3.1, we obtain

$$P_{11}(0,t) = 0.0097e^{\gamma_1 t} + 0.9903e^{\gamma_4 t},$$
$$P_{21}(0,t) = -0.098e^{\gamma_1 t} + 0.098e^{\gamma_4 t},$$
$$P_{11}(1,t) = (0.0188+0.0183t)e^{\gamma_1 t} + (-0.0188+0.9009t)e^{\gamma_4 t},$$
$$P_{21}(1,t) = -(0.0943+0.1853t)e^{\gamma_1 t} + (0.0943+0.0893t)e^{\gamma_4 t},$$
$$P_{11}(2,t) = (-0.0271-0.0195t+0.0173t^2)e^{\gamma_1 t} + (0.0271-0.0078t+0.4098t^2)e^{\gamma_4 t},$$
$$P_{21}(2,t) = -(0.0939+0.1828t+0.1751t^2)e^{\gamma_1 t} + (0.0939+0.0871t+0.0406t^2)e^{\gamma_4 t},$$

$$P_{12}(0,t) = -0.098e^{\gamma_1 t} + 0.098e^{\gamma_4 t},$$
$$P_{22}(0,t) = 0.99e^{\gamma_1 t} + 0.01e^{\gamma_4 t},$$
$$P_{12}(1,t) = -(0.0946+0.1853t)e^{\gamma_1 t} + (0.0946+0.0892t)e^{\gamma_4, t}$$
$$P_{22}(1,t) = (-0.0189+1.872t)e^{\gamma_1 t} + (0.0189+0.0088t)e^{\gamma_4 t}$$
$$P_{12}(2,t) = -(0.0886+0.1751t+0.1773t^2)e^{\gamma_1 t} + (0.0886+0.0871t+0.0406t^2)e^{\gamma_4 t},$$
$$P_{22}(2,t) = (-0.0273-0.045t+1.7693t^2)e^{\gamma_1 t} + (0.0273+0.0173t+0.004t^2)e^{\gamma_4 t}.$$

Similarly, we obtain from (3.22) that

$$P_{11}(t) = 0.5(1+e^{-0.2t}), \qquad P_{12}(t) = 0.5(1-e^{-0.2t}),$$
$$P_{21}(t) = 0.5(1-e^{-0.2t}), \qquad P_{22}(t) = 0.5(1+e^{-0.2t}).$$

These expressions enable us to evaluate the transition probabilities of the service–start embedded Markov chain. (3.4) now specialized to

$$a_{11}^0 = \int_0^1 P_{11}(0,t)dt = 0.6328, \qquad a_{12}^0 = \int_0^1 P_{12}(0,t)dt = 0.02,$$
$$a_{21}^0 = \int_0^1 P_{21}(0,t)dt = 0.02, \qquad a_{22}^0 = \int_0^1 P_{22}(0,t)dt = 0.4329,$$

$$a_{11}^1 = \int_0^1 P_{11}(1,t)dt = 0.2383, \qquad a_{12}^1 = \int_0^1 P_{12}(1,t)dt = 0.0156,$$
$$a_{21}^1 = \int_0^1 P_{21}(1,t)dt = 0.0156, \qquad a_{22}^1 = \int_0^1 P_{22}(1,t)dt = 0.2827,$$

$$a_{11}^2 = \int_0^1 P_{11}(2,t)dt = 0.0682, \qquad a_{12}^2 = \int_0^1 P_{12}(2,t)dt = 0.0079$$
$$a_{21}^2 = \int_0^1 P_{21}(2,t)dt = 0.0078, \qquad a_{22}^2 = \int_0^1 P_{22}(2,t)dt = 0.1462\ .$$

Similarly, (3.21) leads to

$$a_{11} = \int_0^1 P_{11}(t)dt = 0.9532, \qquad a_{12} = \int_0^1 P_{12}(t)dt = 0.0468,$$
$$a_{21} = \int_0^1 P_{21}(t)dt = 0.0468, \qquad a_{22} = \int_0^1 P_{22}(t)dt = 0.9532.$$

The linear–equation system (3.13) is given by

$$\begin{aligned}
Q_{11} &= 0.1Q_{21} + 0.9,\\
Q_{21} &= (0.1/2.1)Q_{11} + (0.1/2.1)Q_{21},\\
Q_{12} &= 0.1Q_{22},\\
Q_{22} &= (0.1/2.1)Q_{12} + (0.1/2.1)Q_{22} + 1.9/2.1\ ,
\end{aligned}$$

with solution

$$Q_{11} = 0.9045, \qquad Q_{21} = 0.0452,$$
$$Q_{12} = 0.0955, \qquad Q_{22} = 0.9548.$$

In our example, (3.3) is specialized to the following system:

$$\begin{aligned}
\sigma_1^{0'} &= (a_{11}^0Q_{11}+a_{12}^0Q_{21})(\sigma_1^{0'}+\sigma_1^{0'}) + (a_{21}^0Q_{11}+a_{22}^0Q_{21})(\sigma_2^{0'}+\sigma_2^{0'}),\\
\sigma_1^{0} &= a_{11}^1(\sigma_1^{0'}+\sigma_1^{0'}) + a_{21}^1(\sigma_2^{0'}+\sigma_2^{0'}) + a_{11}^0\sigma_1^1 + a_{21}^0\sigma_2^1\\
\sigma_2^{0'} &= (a_{11}^0Q_{12}+a_{12}^0Q_{22})(\sigma_1^{0'}+\sigma_1^{0'}) + (a_{21}^0Q_{12}+a_{22}^0Q_{22})(\sigma_2^{0'}+\sigma_2^{0'}),\\
\sigma_2^{0} &= a_{12}^1(\sigma_1^{0'}+\sigma_1^{0'}) + a_{22}^1(\sigma_2^{0'}+\sigma_2^{0'}) + a_{12}^0\sigma_1^1 + a_{22}^0\sigma_2^1,\\
\sigma_1^{1} &= (a_{11}-a_{11}^0-a_{11}^1)(\sigma_1^{0'}+\sigma_1^{0'}) + (a_{21}-a_{21}^0-a_{21}^1)(\sigma_2^{0'}+\sigma_2^{0'}) + (a_{11}-a_{11}^0)\sigma_1^1 + (a_{21}-a_{21}^0)\sigma_2^1\ ,\\
\sigma_2^{1} &= (a_{12}-a_{12}^0-a_{12}^1)(\sigma_1^{0'}+\sigma_1^{0'}) + (a_{22}-a_{22}^0-a_{22}^1)(\sigma_2^{0'}+\sigma_2^{0'}) + (a_{12}-a_{12}^0)\sigma_1^1 + (a_{22}-a_{22}^0)\sigma_2^1\ ,\\
\sigma_1^{0'}&+\sigma_1^{0'}+\sigma_2^{0'}+\sigma_2^{0'}+\sigma_1^1+\sigma_2^1 = 1.
\end{aligned}$$

By substituting the values of all the coefficients into the above system, we obtain

$$\sigma_1^{0'} = 0.5733(\sigma_1^{0'} + \sigma_1^{0'}) + 0.0377(\sigma_2^{0'} + \sigma_2^{0'}),$$
$$\sigma_1^{0} = 0.2384(\sigma_1^{0'} + \sigma_1^{0'}) + 0.0156(\sigma_2^{0'} + \sigma_2^{0'}) + 0.6327\sigma_1^{1} + 0.02\sigma_2^{1},$$
$$\sigma_2^{0'} = 0.0794(\sigma_1^{0'} + \sigma_1^{0'}) + 0.4152(\sigma_2^{0'} + \sigma_2^{0'}),$$
$$\sigma_2^{0} = 0.0156(\sigma_1^{0'} + \sigma_1^{0'}) + 0.2827(\sigma_2^{0'} + \sigma_2^{0'}) + 0.02\sigma_1^{1} + 0.4329\sigma_2^{1},$$
$$\sigma_1^{1} = 0.082(\sigma_1^{0'} + \sigma_1^{0'}) + 0.0113(\sigma_2^{0'} + \sigma_2^{0'}) + 0.3204\sigma_1^{1} + 0.0269\sigma_2^{1},$$
$$\sigma_2^{1} = 0.0112(\sigma_1^{0'} + \sigma_1^{0'}) + 0.2376(\sigma_2^{0'} + \sigma_2^{0'}) + 0.0268\sigma_1^{1} + 0.5204\sigma_2^{1}.$$
$$\sigma_1^{0'} + \sigma_1^{0'} + \sigma_2^{0'} + \sigma_2^{0'} + \sigma_1^{1} + \sigma_2^{1} = 1$$

Solving this system leads to the following solution

$$\sigma_1^{0'} = 0.2591, \qquad \sigma_2^{0'} = 0.1638,$$
$$\sigma_1^{0} = 0.1921, \qquad \sigma_2^{0} = 0.1625,$$
$$\sigma_1^{1} = 0.0621, \qquad \sigma_2^{1} = 0.1604.$$

We next compute $\theta_m^j(x)$ for m = i, 2; j = 0, 1 and $\theta_1(x)$, $\theta_2(x)$. By (3.9), (3.23) and (3.12), we have

$$\theta_1^0(x) = x\int_0^{1-x}\Big[0.9P_{11}(0,t) + 1.9P_{12}(0,t)\Big]dt + \int_{1-x}^{1}(1-t)\Big[0.9P_{11}(0,t) + 1.9P_{12}(0,t)\Big]dt,$$
$$\theta_2^0(x) = x\int_0^{1-x}\Big[0.9P_{21}(0,t) + 1.9P_{22}(0,t)\Big]dt + \int_{1-x}^{1}(1-t)\Big[0.9P_{21}(0,t) + 1.9P_{22}(0,t)\Big]dt,$$
$$\theta_1^1(x) = x\int_0^{1-x}\Big[0.9P_{11}(1,t) + 1.9P_{12}(1,t)\Big]dt + \int_{1-x}^{1}(1-t)\Big[0.9P_{11}(1,t) + 1.9P_{12}(1,t)\Big]dt,$$
$$\theta_2^1(x) = x\int_0^{1-x}\Big[0.9P_{21}(1,t) + 1.9P_{22}(1,t)\Big]dt + \int_{1-x}^{1}(1-t)\Big[0.9P_{21}(1,t) + 1.9P_{22}(1,t)\Big]dt,$$
$$\theta_1(x) = x\int_0^{1-x}\Big[0.9P_{11}(t) + 1.9P_{12}(t)\Big]dt + \int_{1-x}^{1}(1-t)\Big[0.9P_{11}(t) + 1.9P_{12}(t)\Big]dt,$$
$$\theta_2(x) = x\int_0^{1-x}\Big[0.9P_{21}(t) + 1.9P_{22}(t)\Big]dt + \int_{1-x}^{1}(1-t)\Big[0.9P_{21}(t) + 1.9P_{22}(t)\Big]dt.$$

Substituting expressions of $P_{mn}(j,t)$ for m, n = 1, 2; j = 0, 1 correspondingly into this system and integrating the resulting system, we obtain

$$\theta_1^0(x) = 0.4025 + 0.9999x + 0.0059e^{-\gamma_1 x} - 0.4084e^{-\gamma_2 x},$$
$$\theta_2^0(x) = 0.1001 + 1.0003x - 0.0595e^{-\gamma_1 x} - 0.0406e^{-\gamma_2 x},$$

$$\theta_1^1(x) = 1.1561 + 1.0004x + 0.0276e^{-\gamma_1 x} - 1.1837e^{-\gamma_2 x} - 0.0111xe^{-\gamma_1 x} + 0.3715xe^{-\gamma_2 x},$$
$$\theta_2^1(x) = 0.3824 + 1.0047x - 0.2203e^{-\gamma_1 x} - 0.1621e^{-\gamma_2 x} + 0.1125xe^{-\gamma_1 x} + 0.0385xe^{-\gamma_2 x},$$
$$\theta_1(x) = -10.2341 - 1.1x - 0.7x^2 + 10.2341e^{0.2x},$$
$$\theta_2(x) = 10.2341 + 3.9x - 0.7x^2 - 10.2341e^{0.2x}.$$

From the above expressions, we can easily get

$$\theta_1^0(1) = 0.3472, \qquad \theta_2^0(1) = 0.5487,$$
$$\theta_1^1(1) = 0.0934, \qquad \theta_2^1(1) = 0.2489,$$
$$\theta_1(1) = 0.4659, \qquad \theta_2(1) = 0.9341.$$

Now, we have only two unknowns to be computed, i.e. α^3, the blocking probability and E(N), the average number of customers arrivals in a service period. From (3.26), we have

$$\beta_1 = (\sigma_1^{0'} + \sigma_1^0)\Big[\theta_1(1) - \theta_1^0(1) - \theta_1^1(1)\Big] + \sigma_1^1\Big[\theta_1(1) - \theta_1^0(1)\Big],$$
$$\beta_2 = (\sigma_2^{0'} + \sigma_2^0)\Big[\theta_2(1) - \theta_2^0(1) - \theta_2^1(1)\Big] + \sigma_2^1\Big[\theta_2(1) - \theta_2^0(1)\Big].$$

Or equivalently,

$$\beta_1 = 0.0188, \qquad \beta_2 = 0.1064.$$

Therefore (by (3.11a))

$$\alpha^3(1) = \frac{(1-\sigma_1^{0'}-\sigma_2^{0'})(\beta_1+\beta_2)}{E(N) - (\sigma_1^{0'}+\sigma_2^{0'})(\beta_1+\beta_2)}$$
$$= \frac{0.0723}{E(N) - 0.0529}.$$

In our case, (3.29) specializes to

$$E(N) = \Big[(\sigma_1^{0'}+\sigma_1^0+\sigma_1^1)\theta_1(1) + (\sigma_2^{0'}+\sigma_2^0+\sigma_2^1)\theta_2(x)\Big] = 0.6938.$$

Therefore,

$$\alpha^3(1) = \frac{0.0723}{E(N) - 0.0529} = 0.1128 .$$

Thus, our final solution is expressed as

$$\alpha^0 = 0.3752,$$
$$\alpha^1(x) = 0.4063\theta_1^0(x) + 0.2939\theta_2^0(x),$$
$$\alpha^2(x) = 0.4063\theta_1^1(x) + 0.2939\theta_2^1(x) + 0.0559\theta_1^0(x) + 0.1445\theta_2^0(x),$$
$$\alpha^3(x) = 0.4623[\theta_1(x) - \theta_1^0(x)] + 0.4383[\theta_2(x) - \theta_2^0(x)] - 0.4063\theta_1^1(x) - 0.2939\theta_2^1(x),$$

That is,

$$\alpha^0 = 0.3752,$$
$$\alpha^1(x) = 0.193 + 0.7003x - 0.0151e^{-\gamma_1 x} - 0.1779e^{-\gamma_2 x},$$
$$\alpha^2(x) = 0.6191 + 0.9022x - 0.0618e^{-\gamma_1 x} - 0.5573e^{-\gamma_2 x} + \\ + 0.0316xe^{-\gamma_1 x} + 0.1622xe^{-\gamma_2 x},$$
$$\alpha^3(x) = -1.058 - 0.4017x - 0.6304x^2 + 0.2459e^{0.2x} + \\ + 0.0768e^{-\gamma_1 x} + 0.7352e^{-\gamma_2 x} - 0.0285xe^{-\gamma_1 x} - 0.1622xe^{-\gamma_2 x}.$$

3.3.4 Computational Scheme for Specialized Service Times

We have now completed the analysis of α^0 and $\alpha^j(x)$ for the C/G/1/K queue. As we have seen in the previous sections that the computation of α^0 and $\alpha^j(x)$ for the C/G/1/K queue requires the solutions of systems of differential equations of the forms given in (3.16) and (3.22). In addition, numerical integrations over infinite intervals are also required for the computation of the important quantities a_{mr}^j and $\theta_m^j(x)$ for m, r $\in \mathbb{N}$ and $0 \le j \le K-1$, given in (3.4) and (3.9) respectively. From the numerical–computation point of view these calculations involve significantly more computational effort than simply working with linear equations. When the distribution of service times in the system is

assumed to be arbitrary, this is a price we have to pay. However, we can avoid such computational burden when the service–time distribution in the C/G/1/K queue system is either generalized hyperexponential or Erlang. In these special cases, we only need to solve systems of linear equations. Hence, the procedure will be much efficient. Since any distribution can be approximated by a generalized hyperexponential distribution our method to be introduced in this section has substantial value in applications. We begin by defining the Laplace transforms of $P_{ij}(n,t)$ and $P_{ij}(t)$ as

$$\mathscr{P}_{ij}(n,s) = \int_0^\infty P_{ij}(n,t)e^{-st}dt, \tag{3.31}$$

and

$$\mathscr{P}_{ij}(s) = \int_0^\infty P_{ij}(t)e^{-st}dt, \tag{3.32}$$

respectively. We now assume that the service–time distribution $G(\cdot)$ has the form

$$G(x) = \sum_{k=1}^{L} \xi_k(1 - e^{-\mu_k x}), \tag{3.33}$$

where L is a finite positive integer and the "weights" ξ_k for $1 \le k \le L$ satisfy the condition $\Sigma_k \xi_k = 1$.

From (3.4), we then have that for m, $r \in \mathbb{N}$ and $0 \le j \le K-1$,

$$\begin{aligned} a_{mr}^{j} &= \int_0^\infty P_{mr}(j,t)\, dG(t) \\ &= \int_0^\infty P_{mr}(j,t)\, d\left[\Sigma_k \xi_k(1 - e^{-\mu_k t})\right] \\ &= \Sigma_k \xi_k \mu_k \int_0^\infty P_{mr}(j,t)e^{-\mu_k x}dt \\ &= \Sigma_k \xi_k \mu_k \mathscr{P}_{mr}(j,\mu_k), \end{aligned} \tag{3.34}$$

where the last equality in (3.34) is a consequence of (3.32). Similarly, we have for $m, r \in \mathbb{N}$

$$a_{mr} = \int_0^\infty P_{mr}(t)dG(t)$$

$$= \Sigma_k \xi_k \mu_k \mathscr{P}_{mr}(\mu_k). \tag{3.35}$$

(3.34) and (3.35) show that the integrals in the expression of a_{mr}^j and a_{mr} can be expressed as a weighted sum of the Laplace transform of $P_{mr}(j,t)$ and $P_{mr}(t)$ respectively. When these expressions are used as final formulas for numerical computation (see Step 2 of our computational procedure, Section 3.3.2), the corresponding numerical integrations will no longer be necessary.

We can also express integral formulas for $\theta_m^j(x)$ as weighted sums of $\mathscr{P}_{mr}(j,\mu_k)$. From (3.9) and (3.12), we have that for $m \in \mathbb{N}$ and $0 \le j \le K-1$,

$$\theta_m^j(x) = \int_0^\infty [G(t+x) - G(t)]dA_m^{j+1}(t)$$

$$= \int_0^\infty [G(t+x) - G(t)]\Sigma_r P_{mr}(j,t)\lambda_r dt$$

$$= \int_0^\infty \Sigma_k \xi_k \left[e^{-\mu_k t} - e^{-\mu_k (t+x)}\right] \Sigma_r P_{mr}(j,t)\lambda_r dt.$$

By changing the order of summation and integration in the last equality, we now obtain

$$\theta_m^j(x) = \Sigma_k \left[\xi_k (1 - e^{-\mu_k x}) \int_0^\infty e^{-\mu_k t} \Sigma_r P_{mr}(j,t)\lambda_r dt\right]$$

$$= \Sigma_k \left[\xi_k (1 - e^{-\mu_k x}) \Sigma_r \lambda_r \mathscr{P}_{mr}(j,\mu_k)\right], \tag{3.36}$$

a linear combination of $\mathscr{P}_{mr}(j,\mu_k)$. Similarly, we have, by (3.23) that

$$\theta_m^j(x) = \int_0^\infty [G(t+x) - G(t)]\Sigma_r P_{mr}(t)\lambda_r dt$$

$$= \Sigma_k \left[\xi_k (1 - e^{-\mu_k x}) \Sigma_r \lambda_r \mathcal{P}_{mr}(\mu_k) \right]. \tag{3.37}$$

Now, all integrals in our computational scheme (Section 3.3.2) have been changed into linear combinations of $\mathcal{P}_{mr}(j,\mu_k)$ and $\mathcal{P}_{mr}(\mu_k)$. This is an important simplification of our numerical scheme, because linear–combination schemes are implemented easily in numerical computation, while integrations, on the other hand, are implemented through various approximation schemes that require large volume of operations to get reasonably accurate results, and hence are much more time consuming (especially for integrations over infinite intervals).

The Laplace transforms of $P_{mr}(j,t)$ and $P_{mr}(t)$, on the other hand, satisfy two systems of linear equations respectively. These two systems of linear equations can be derived by taking the corresponding Laplace transforms in equations (3.16) and (3.22) respectively. From (3.16), we obtain

$$s\mathcal{P}_{mr}(j,s) - P_{mr}(j,0) = \nu_m \Sigma_i p_{mi} \mathcal{P}_{ir}(j,s) - (\lambda_m + \nu_m)\mathcal{P}_{mr}(j,s) + \lambda_m \mathcal{P}_{mr}(j-1,s),$$

where we have substituted the relation

$$\int_0^\infty e^{-st} \left[\frac{d}{dt} P_{mr}(j,t) \right] dt = s\mathcal{P}_{mr}(j,s) - P_{mr}(j,0)$$

to obtain the left–hand side terms. By using boundary condition (3.14), we obtain, for $m, r \in \mathbb{N}$ and $0 \le j \le K-1$ that

$$s\mathcal{P}_{mr}(j,s) - \delta_{j0}\delta_{mr} = \nu_m \Sigma_i p_{mi} \mathcal{P}_{ir}(j,s) - (\lambda_m + \nu_m)\mathcal{P}_{mr}(j,s) + \lambda_m \mathcal{P}_{mr}(j-1,s). \tag{3.38}$$

Similarly, using boundary condition $P_{ij}(0) = \delta_{ij}$, we obtain from (3.22) that

$$s\mathscr{P}_{ij}(s) - \delta_{ij} = \Sigma_m \nu_i p_{im} \mathscr{P}_{mj}(s) - \nu_i \mathscr{P}_{ij}(s), \qquad \text{for } i, j \in \mathbb{N}. \tag{3.39}$$

In contrast with (3.16) and (3.22) which define two system of differential equations, (3.38) and (3.39) are two systems of linear equations with parameter s. And therefore, (3.38) and (3.39) are much easier to solve. One possible procedure for solving (3.38) and (3.39) is as follows: solve (3.38) and (3.39) for the analytic solutions of $\mathscr{P}_{mr}(j,s)$ and $\mathscr{P}_{ij}(s)$, leaving s as a parameter; then let this parameter in $\mathscr{P}_{mr}(j,s)$ and $\mathscr{P}_{ij}(s)$ take specific values of μ_k for k = 1, 2, ..., L to obtain $\mathscr{P}_{mr}(j,\mu_k)$ and $\mathscr{P}_{ij}(\mu_k)$ for m, r $\in \mathbb{N}$. We do not, however, recommend this scheme, because numerical computation generally does not produce parametric solutions. In practice, we should replace s in (3.38) and (3.39) by specific values of μ_k before solving for them. Consequently, in each of (3.38) and (3.39), there are L systems of linear equations that need to be solved; and in each of (3.16) and (3.22), there are $K \times |\mathbb{N}| + 1$ or $|\mathbb{N}| + 1$ systems of linear equations that need to be solved (see Section 3.3.1 and Section 3.3.2). When L is comparatively small, this computational scheme will be very efficient, in addition to the important advantage of taking no numerical integrations in the procedure.

The same argument applies when the service–time distribution is Erlang. The only difference is that a_{mr}^{j} and $\theta_m^j(x)$ are no longer linear combinations of $\mathscr{P}_{mr}(j,s)$. Instead, they are expressed in terms of the derivatives of $\mathscr{P}_{mr}(j,s)$. Let G now take the form

$$G(x) = \int_0^x e^{-\mu y} \frac{(\mu y)^{k-1}}{k!} \mu \, dy, \tag{3.40}$$

which is the Erlang distributed with parameters (μ,k) We have, from (3.4), that

$$a_{mr}^{j} = \int_0^\infty P_{mr}(j,t) dG(t)$$

$$= \int_0^\infty P_{mr}(j,t) \, e^{-\mu t} \frac{(\mu t)^{k-1}}{(k-1)!} \mu \, dt$$

$$= \frac{\mu^k}{(k-1)!} \int_0^\infty P_{mr}(j,t)\, e^{-\mu t} t^{k-1} dt.$$

Applying the transform relation

$$\int_0^\infty e^{-\mu t} \left[P_{mr}(j,t) t^k \right] dt = (-1)^k \frac{d^k}{d\mu^k} \mathscr{P}_{mr}(j,\mu), \tag{3.41}$$

we obtain, for $m, r \in \mathbb{N}$ and $0 \le j \le K-1$, that

$$a_{mr}^j = \frac{\mu^k}{(k-1)!} (-1)^{k-1} \frac{d^{k-1}}{d\mu^{k-1}} \mathscr{P}_{mr}(j,\mu). \tag{3.42}$$

Similarly, a_{mr} is expressed as

$$a_{mr}^j = \frac{\mu^k}{(k-1)!} (-1)^{k-1} \frac{d^{k-1}}{d\mu^{k-1}} \mathscr{P}_{mr}(\mu). \tag{3.43}$$

We also have, by (3.9) and (3.12), that

$$\theta_m^j(x) = \int_0^\infty [G(t+x) - G(t)] dA_m(j+1,t)$$

$$= \int_0^\infty \left[\int_t^{t+x} e^{-\mu y} \frac{(\mu y)^{k-1}}{(k-1)!} \mu dy \right] \Sigma_r P_{mr}(j,t) \lambda_r dt$$

$$= \frac{\mu^k}{(k-1)!} \int_0^\infty \left[\int_t^{t+x} e^{-\mu y} y^{k-1} dy \right] \Sigma_r \lambda_r P_{mr}(j,t) dt.$$

Taking derivative of $\theta_m^j(x)$ with respect to x in the last equation, we obtain

$$\frac{d}{dx}\theta_m^j(x) = \frac{\mu^k}{(k-1)!} \int_0^\infty \left\{ e^{-\mu(t+x)} (t+x)^{k-1} \Sigma_r \lambda_r P_{mr}(j,t) \right\} dt$$

$$= \frac{\mu^k}{(k-1)!} \int_0^\infty \left\{ e^{-\mu t} e^{-\mu x} \left[\sum_{i=0}^{k-1} \binom{k-1}{i} t^i x^{k-i-1} \right] \Sigma_r \lambda_r P_{mr}(j,t) \right\} dt$$

$$= \frac{\mu^k}{(k-1)!}\sum_{i=0}^{k-1}\left\{\left[\binom{k-1}{i}x^{k-i-1}e^{-\mu x}\right]\left[\Sigma_r\lambda_r\int_0^\infty e^{-\mu t}P_{mr}(j,t)t^i dt\right]\right\},$$

where we used the binomial expansion of $(t+x)^{k-1}$ to obtain the second equality. (The interchange of the order of the derivative and the integration can be easily justified.) After substituting (3.41), this last expression further reduces to

$$\frac{d}{dx}\theta_m^j(x) = \frac{\mu^k}{(k-1)!}\sum_{i=0}^{k-1}(-1)^i\left\{\left[\binom{k-1}{i}x^{k-i-1}e^{-\mu x}\right]\left[\Sigma_r\lambda_r\frac{d^i}{d\mu^i}\mathcal{P}_{mr}(j,\mu)\right]\right\}. \tag{3.44}$$

Integrating (3.44) over the interval [0, x] and using the condition $\theta_m^j(0) \equiv 0$ and the relation

$$\int_0^x t^n e^{-\mu t}dt = \frac{n!}{\mu^{n+1}}\left[1 - \sum_{r=0}^{n}\frac{(\mu x)^r}{r!}e^{-\mu x}\right],$$

we finally obtain

$$\theta_m^j(x) = \frac{\mu^k}{(k-1)!}\sum_{i=0}^{k-1}(-1)^i\left\{\binom{k-1}{i}\frac{(k-i-1)!}{\mu^{k-i}}\left[1 - \sum_{n=0}^{k-i-1}\frac{(\mu x)^n}{n!}e^{-\mu x}\right]\left[\Sigma_r\lambda_r\frac{d^i}{d\mu^i}\mathcal{P}_{mr}(j,\mu)\right]\right\}$$

$$= \sum_{i=0}^{k-1}\left\{\frac{(-\mu)^i}{i!}\left[1 - \sum_{n=0}^{k-i-1}\frac{(\mu x)^n}{n!}e^{-\mu x}\right]\left[\Sigma_r\lambda_r\frac{d^i}{d\mu^i}\mathcal{P}_{mr}(j,\mu)\right]\right\}. \tag{3.45}$$

Similarly, we have

$$\theta_m(x) = \sum_{i=0}^{k-1}\left\{\frac{(-\mu)^i}{i!}\left[1 - \sum_{n=0}^{k-i-1}\frac{(\mu x)^n}{n!}e^{-\mu x}\right]\left[\Sigma_r\lambda_r\frac{d^i}{d\mu^i}\mathcal{P}_{mr}(\mu)\right]\right\}. \tag{3.46}$$

In (3.42), (3.43), (3.45), and (3.46), calculations of a_{mr}^j and $\theta_m^j(x)$ require the derivatives of $\mathcal{P}_{mr}(j,\mu)$ and $\mathcal{P}_{mr}(\mu)$ up to the $(k-1)^{th}$ order. Similar to the generalized–hyperexponential case, these derivatives can be derived by solving k systems of linear equations. Take, for example, the derivatives of $\mathcal{P}_{mr}(j,\mu)$. We first solve $\mathcal{P}_{mr}(j,\mu)$ from (3.38), and then take derivatives of (3.38) with respect to t and solve the resulting system of equations for $\frac{d}{d\mu}\mathcal{P}_{mr}(j,\mu)$, after substituting the obtained $\mathcal{P}_{mr}(j,\mu)$ into the system. This process is then repeated until all of the $\frac{d^{k-1}}{d\mu^{k-1}}\mathcal{P}_{mr}(j,\mu)$ are solved. In this step, k systems of linear equations need to be solved *consecutively.* It is interesting to note that in the

generalized–hyperexponential case, we also need to solve systems of linear equations for $\mathscr{P}_{mr}(j,\mu_k)$, $k \in \{1, 2, ..., L\}$. But the operations for the generalized–hyperexponential–service case are carried out in a parallel manner (while in the Erlang–service case they are done consecutively), reflecting the structural difference between Erlang distribution and the generalized hyperexponential distribution.

Finally, we point out that the method of analysis and the computational procedures discussed in this chapter also apply when the service times are not renewal. All that is needed is sufficient information about the relationship between consecutive service times so that a suitable service–start Markov chain can be formulated. The only extra effort needed is to include necessary additional parameters in the state space of the service–start Markov chain to keep track of the information about the service process. For specific examples (semi–Markovian service, exceptional first services, and server vacations in the M/G/1/K model), please see Niu and Cooper [1989].

REFERENCES

Avi–Itzhak, B., Maxwell, W.L., and Miller, L.W. [1965]. "Queueing with Alternating Priorities." *Operations research*, No. 13, 306–318.

Balachandran, K. R. [1973], "Control Policies for a Single Server System." *Management Science*, Vol. 19, No. 9, 1013–1018.

Balachandran, K. R., Tijms, H. [1975], "On the D–Policy for the M/G/1 Queue." *Management Science*, Vol. 21, No. 9, 1073–1076.

Boxma, O. J. [1976], "Note on a Control Problem of Balachandran And Tijms." *Management Science*, Vol. 22, No. 8, 916–917.

Brauer, Fred, Nohel, John A., Schneider, Hans, [1970] *Linear Mathematics: An Introduction to Linear Algebra and Linear Differential Equations.* W.A. Benjamin, Inc, New York.

Cohen, J. W. [1982], *The Single Server Queue*, Second Edition. North–Holland, Amsterdam. First Edition, 1969, American Elsevier, New York.

Cooper, R. B. [1981], *Introduction to Queueing Theory*, Second Edition. North–Holland (Elsevier). First Edition, 1972, Macmillan.

Cooper, R. B. and Murray, G. [1969], "Queues Served in Cyclic Order." *The Bell System Technical Journal* 48, No. 3 (March) 675–689.

Crabill, T., D. Gross and N. Magazine [1977], "A Classified Bibliography of Research on Optimal Design and Control of Queues." *Operations Research*, Vol. 25,

Doshi, B. T. [1985], "A Note on Stochastic Decomposition in a GI/G/1 Queue with vacation or Set–up Times." *Journal of Applied Probability*, Vol. 22, No. 2, 419–428.

Feller, W. [1968], *An Introduction to Probability Theory And Its Applications*, Vol. 2, Second Edition, John Wiley & Sons.

Gaver, D.P. [1962]. "A Waiting Line with Interrupted Service, Including Priorities." *Proceedings of the Royal Statistical Society*, Seeries B, Vol. 24, 73–90.

Heyman, D. P. [1977], "The T–Policy for the M/G/1 Queue." *Management Science*, Vol. 23, No 7, 775–778.

Keilson, J. [1962], "Queues Subject to Service Interruption." *Annals of Mathematical Statistics*, Vol. 33, 1314–1322.

Keilson, J. [1966], "The Ergodic Queue Length Distribution for Queueing Systems with Finite Capacity." *Journal of the Royal Statistical Society*, B 28, 190–201.

Kella, O. [1986], "The Threshold Policy in the M/G/1 Queue with Server Vacations." *Tech. Report, Yale School of Organization and Management, New Haven*, CT.

Lee, H.S. and M.M. Srinivasan [1987], "Control policies for the $M^X/G/1$ Queueing Systems." *Tech. Report 87–20, Department of Industrial and Operations Engineering. The University of Michigan, Ann Arbor.*

Lee, H.S. and M.M. Srinivasan [1989], "Control policies for the $M^X/G/1$ Queueing Systems." *Management Science*, Vol. 35, No. 6, 708–721.

Neuts, M.F. [1979], "A Versatile Markovian Point Process." *Journal of Applied Probability*, Vol. 16, No. 4, pp.764–779.

Neuts, M.F. [1981], *Matrix–Geometric Solutions in Stochastic Models: An Algorithmic Approach*, The Johns Hopkins University Press, Baltimore.

Niu, S.–C. [1988], "Representing Workloads in GI/G/1 Queues Through the Preemptive–Resume LIFO Queue Discipline." *Queueing Systems*, No. 3, 157–178.

Niu, S.–C. and Cooper, R. B. [1989]. "Transform–Free Results for M/G/1 Finite and Infinite Capacity Queues, with Generalizations." (in preparation for pub

Ramaswami, V. [1980], "The N/G/1 Queue and Its Detailed Analysis." *Advances in Applied Probability*, Vol. 12, No. 1, 222–261.

Regterschot, G.J.K. and de Smit, J.H.A. [1986], "The Queue M/G/1 with Markov Modulated Arrivals and Services." *Mathematics of Operations Research*, vol. 11, No. 3, 465–483.

Ross, S. M. [1983], *Stochastic Processes.* Wiley, New York.

Takács, L. [1963]. "Delay Distributions for One Line with Poisson Input, General Holding Times, and Various Orders of Service." *The Bell System Technical Journal*, Vol. 43, 487–453.

Skinner, C.E. [1967]. "A Priority Queueing System with Server–Walking Time." *Operations Research*, Vol. 15, No. 2, 278–285.

Welch, P. D. [1964], "On a Generalized M/G/1 Queueing Process in Which the First Customer of Each Busy Period Receives Exceptional Service." *Operations Research*, Vol. 12, 736–752.

Wishart, D. M. G. [1961], "An Application of Ergodic Theorems in the Theory of Queues." *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 2, University of California Press, 581–592.

Wolff, R. W. [1982], "Poisson Arrivals See Time Averages." *Operations Research*, Vol. 30, 223–231.

Yadin, M., Noar, P. [1963], "Queueing Systems with a Removable Service Station." *Operational Research Quarterly*, Vol. 14, No. 4, 393–405.

VITA

Jingwen Li was born on December 13, 1957 in Yanting, Sichuan, People's Republic of China, to Tongxin Li and Yunhua Se. After graduating from the South–Hill High School, Mianyang, Sichuan, China, during the "cultural revolution", he was sent in 1974, for the the so–called "re–education" to the countryside in the Eastern Village, Houma, Shanxi, China, where he worked two and half years before he was employed by the Linfen Tractor factory in Linfen, Shanxi, China. In 1978, he joined the Nanjing Institute of Meteorology, Nanjing, Jiangsu, China, where he earned the degree of Bachelor of Science with a major in synoptic and dynamic weather forecasting. He was a recipient of the National Merit Scholarship for Overseas Studies, sponsored by the Education Ministry of the People's Republic of China. From September 1984 to August 1985, he studied at the Colorado State University, Fort Collins, Colorado, U.S.A., and received the degree of Master of Science with a major in the atmospheric dynamics, in August 1985. Since then, he has been working on a Ph.D. degree in management sciences with a specialization in operations research in the School of Management, The University of Texas at Dallas, Richardson, Texas, U.S.A.