UNRELIABLE RETRIAL QUEUES IN A RANDOM ENVIRONMENT

DISSERTATION

James D. Cordeiro, Jr., Major, USAF

AFIT/DS/ENS/07-03

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

Research sponsored by the Air Force Office of Scientific Research. The United States Government is authorized to reproduce and distribute reprints notwithstanding any copyright notation thereon. The views and conclusions contained in this dissertation are those of the author and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Office of Scientific Research or the United States Government.

AFIT/DS/ENS/07-03

UNRELIABLE RETRIAL QUEUES IN A RANDOM ENVIRONMENT

DISSERTATION

Presented to the Faculty
Department of Operational Sciences
Graduate School of Engineering and Management
Air Force Institute of Technology
Air University
Air Education and Training Command
in Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy

James D. Cordeiro, Jr., A.B., M.S., M.S.
Major, USAF

September 2007

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

AFIT/DS/ENS/07-03

UNRELIABLE RETRIAL QUEUES IN A RANDOM ENVIRONMENT

James D. Cordeiro, Jr., A.B., M.S., M.S.

Major, USAF

Approved:

Dr. Jeffrey P. Kharoufeh Committee Chair	Date
Dr. Adedeji B. Badiru Dean's Representative	Date
Dr. Sharif H. Melouk Committee Member	Date
Dr. Mark E. Oxley Committee Member	Date

Accepted:

M.U. Thomas Dean, Graduate School of Engineering and Management	Date

Table of Contents

Page

List of Figures

List of Tables

AFIT/DS/ENS/07-03

Abstract

This dissertation investigates stability conditions and approximate steady-state performance measures for unreliable, single-server retrial queues operating in a randomly evolving environment. In such systems, arriving customers that find the server busy or failed join a retrial queue from which they attempt to regain access to the server at random intervals. Such models are useful for the performance evaluation of communications and computer networks which are characterized by time-varying arrival, service and failure rates. To model this time-varying behavior, we study systems whose parameters are modulated by a finite Markov process. Two distinct cases are analyzed. The first considers systems with Markov-modulated arrival, service, retrial, failure and repair rates assuming all interevent and service times are exponentially distributed. The joint process of the orbit size, environment state, and server status is shown to be a tri-layered, level-dependent quasi-birth-and-death (LDQBD) process, and we provide a necessary and sufficient condition for the positive recurrence of LDQBDs using classical techniques. Moreover, we apply efficient numerical algorithms, designed to exploit the matrix-geometric structure of the model, to compute the approximate steady-state orbit size distribution and mean congestion and delay measures. The second case assumes that customers bring generally distributed service requirements while all other processes are identical to the first case. We show that the joint process of orbit size, environment state and server status is a level-dependent, M/G/1-type stochastic process. By employing regenerative theory, and exploiting the M/G/1-type structure, we derive a necessary and sufficient condition for stability of the system. Finally, for the exponential model, we illustrate how the main results may be used to simultaneously select arrival and service rates that minimize the mean time customers spend in orbit, subject to bound and stability constraints.

Acknowledgements

Looking back at the expanse of time it has taken me to reach this point in my life and my academic career, I am reminded of the very special people who have contributed to this achievement. I would like to thank my parents for stressing the value of education and sacrificing their wellbeing in order to provide it to me. They are living proof that a parent's love knows no bounds, and I am forever grateful for the gifts that they have given me.

To my wife and our two children, I give my thanks for the patience and understanding that they have shown me despite the long hours that I have spent away from home working on this dissertation. My wife has had to deal with the day-to-day affairs that I have not been able to mind, while our children have endured my absence during some important moments in their lives. For this I sincerely apologize and thank them for being the wonderful family that they have always been.

Finally, to my advisor, Dr. Jeff Kharoufeh, I extend my most sincere thanks for all the hardships *he* has had to endure on my behalf. There is probably no other professor who cares for his students' well-being and for the integrity of research conducted under his supervision. I would have languished in a perpetual state of angst with my intended academic path if he had not convinced me to pursue my interests and talents rather than meet others' expectations. There are teachers and mentors whom you will always remember as a prime influence in your life, and he will figure most prominently on this list. I am also grateful for the other members of my research committee, namely Drs. Sharif Melouk and Mark Oxley, for their careful reading of this document and helpful suggestions, and Dr. Deji Badiru who served as the Dean's Representative on the committee.

James D. Cordeiro, Jr.

UNRELIABLE RETRIAL QUEUES IN A
RANDOM ENVIRONMENT

1. Introduction

Since the introduction of the first digital computers some 70 years ago, networks of information systems have become an integral part of the world's financial, industrial, educational, and governmental institutions. We have come to depend almost solely on these information systems and the networks upon which they reside for managing our financial systems, utility infrastructures, and countless other functions essential to modern living. The negative side of this transformation is that these systems are subject to the unquestionably hit-or-miss task of relaying data through Internet pathways that are scarcely controlled. Even the internal networks of large and small organizations are not immune from the often critical risk of losing the ability to exchange information in an accurate and timely fashion. For this reason, it has become necessary to understand the factors that influence the degradation, and ultimately the failure, of information systems.

1.1 Background and Motivation

Maintaining the integrity of information networks has become a universal concern for information technology managers from all areas of society. Corporations and smaller businesses that rely on the internet for sales, marketing, or other direct support must not only be able to communicate with their clients and suppliers, but they must do so in a *prompt manner*, all the while sending and receiving large quantities of information over global distances. Governmental organizations find common ground with the private sector in their desire to maintain and improve information flow. For example, timely communication is an essential staple in military applications. Battle effectiveness is no longer measured by sheer volumes of materiel and manpower, but rather by precision strikes, which themselves are enabled by correct intelligence and the effective coordination of battlefield commanders distributed over wide geographic areas. The information systems network has become, in fact, the single most important center of gravity of opposing military forces, and not just in conflict, but particularly for private entities and even nations in conducting their day-to-day business.

Since computer networks are subject to seemingly-random phenomena over very small time scales, they are difficult to observe directly. It is therefore advantageous to employ mathematical models as proxies for empirical observations or direct experimentation. One such model that is frequently used by analysts in the computer and telecommunications industries is the *queueing system.* A queue consists of an input process for arrivals, a waiting room (otherwise known as a *buffer*), and *servers* that process the arrivals. The input is itself a random process that describes the timing of arrivals to the queue, and is also characterized by the behavior of these arrivals; for instance, these may arrive individually or in groups (*batches*), or they may choose to renege and abandon the queue entirely. The service mechanism of a queueing system is likewise determined by a random process, but one that describes the duration of service episodes. It is furthermore regulated by a *queueing discipline* which dictates the order in which arriving entities are processed. Either or both of the input and service processes may be deterministic, as in an automated production line, or probabilistic, as in a traffic flow model. In the classical queueing models, servers are usually assumed to be reliable. Therefore, entities that leave the system do so only because they have received service, and then they depart permanently. However, it is apparent that the assumption of failure-free operation is not a realistic one, and so the classical system fails to incorporate an essential stochastic characteristic: server failure. Understanding systems in which such failures occur is particularly important to those who analyze communication networks.

It may be unrealistic to expect entities to be lost by a system even if the server is busy or has failed. Current data-link protocols, for instance, usually employ a handshake procedure to verify that a successful communication has been made, during which time the sent data is stored in a buffer for later re-transmission if necessary. Such a protocol is described in [110] and operates in the following manner. The sending device starts off the process by transmitting an information packet to the receiving device. It is important to note here that it *keeps the packet* until it has received verification from the receiver that it has arrived uncorrupted, a task which is accomplished via the *checksum* method.[1] If the receiver determines that it has received the packet correctly, then it sends an acknowledgment (ACK) back to the sending device, together with bits that will allow the sender to verify the integrity of the ACK. If, in the course of this exchange, the packet or the ACK becomes corrupted, then the sender will determine that it did not receive a valid ACK. It will then wait for a prescribed *timeout* period in order to determine whether or not a transmission error has occurred and resend the packet. In the parlance of queueing theory, such a mechanism in which ejected (or rejected) customers return at random intervals until they receive service is called a *retrial queue.* Retrial queues have application in a wide variety of fields, and, as this example clearly shows, they are particularly useful in describing communication systems.

A retrial queue is similar to an ordinary queueing system in that there is an arrival process and one or more servers. The fundamental differences are that (i) entities who enter during a down or busy period of the server or servers may reattempt service at some random time in the future, and (ii) a waiting room, which

[1] The checksum method is used for the detection of errors in transmitted messages. A computation is performed on the essential bits of a message, stored, and then transmitted along with the message. The same computation is performed on the receiving end, and a comparison is made to the transmitted checksum information.

is known as a *primary queue* in the context of retrial queues, is not mandatory. In place of the ordinary waiting room is a buffer called an *orbit* to which entities proceed after an unsuccessful attempt at service, and from which they retry service according to a given probabilistic or deterministic policy. Note that the orbit may or may not be capacitated depending upon the application. More likely than not, orbits that simply represent a pool of returning customers rather than an actual physical waiting area will have an unbounded capacity. A typical example of an infinite-capacity retrial queue orbit would be the pool of customers dialing into a call center who, having abandoned their service request after waiting (i.e. left the primary queue), try again at a later time (i.e. join the retrial orbit).

A key distinguishing feature of retrial queues is the way in which retrials are conducted. Customers may reattempt service independently of each other at random or fixed intervals of time, or they may follow more structured policies. For instance, one might impose a queueing discipline such as 'first-in-first-out' (FIFO) or 'last-in-first-out' (LIFO) upon the entities in the orbit. Yet another policy is to classify arrivals to the retrial queue and establish a prioritization scheme based upon class membership. Some systems may also incorporate *impatient* customers who will retry a finite number of times or for a fixed time interval[2] (which could have zero duration) and then leave the system permanently. Any combination of these priorities may be imposed as well, but the analyst must keep in mind the analytical complexity that these features may introduce into the system description and the numerical computation of the system measures.

A classic example of a retrial queue application may be found in the description of caller behavior in cellular telephone networks. In the basic model, a cellular service area is partitioned into subareas known as cells, each of which contains a base station

[2] The time interval may vary randomly for each customer or be a deterministic value that is assigned by class or some other criteria.

that in turn patches each call that it handles into the national phone network.[3] Calls that originate in one cell but move into an adjoining cell are assigned a certain frequency in the new cell should one exist. If one does not exist, then the call is dropped. The underlying assumption of the retrial model is that the caller will try again and then leave the system when the conversation is terminated. Extensions of this model may, for example, include impatient or prioritized callers as well (consider 911). Regardless, it is not unreasonable to assume that persons who initiate calls do so independently of each other.

The net effect of the partitioning of a service area in a cellular network is an increase in system call capacity that ultimately derives from the reuse of frequencies amongst the various cells. Borst, *et al.* [23] describe a modified version of the original cellular network architechture called a *layered network* that further augments network capacity. In this arrangement, cells are themselves partitioned into what are termed *microcells.* These sub-partitions are optimally used to cover small areas that experience a high volume of calls, which, in the jargon of the cellular communications industry, are known as *hotspots.* Calls that are blocked at the microcell level are sent to overflow buffers at the macrocell level. Here the call waits until it is either assigned a channel by the macrocell or a channel opens up in the original microcell (*repacking*). Retrials correspond here to the repacking process, which occurs according to constant exponential retrial rates that depend upon the overflow buffer. Such a scheme tremendously increases the capacity of each cell since (i) channels may be reused more frequently, and in places where they are most needed, and (ii) the repacking scheme frees up additional channels in the macrocell for use elsewhere.

Retrial queues are likewise prevalent in the evaluation and design of computer networks as they are in telecommunications. Libman and Orda [72] describe an optimal scheme for connections to internet websites from client computers. One of the

[3] In the telecommunications jargon, this is often referred to as 'POTS', which is an acronym for 'plain old telephone system'.

many aggravating issues for users of the internet is the problem of excessively long wait times when attempting connections to slow or inoperative web servers. This wait duration, known formally as the *timeout* duration, is defined as the maximum time that a server allows between the receipt of a connection request and the acknowledgement by the server of failure or success. Current protocols set a conservatively large timeout in order to guarantee a high probability of receipt of acknowledgement. This policy works reasonably well in systems with a low blocking probability, but otherwise creates unnecessarily lengthy waiting periods before a 'failure-to-connect' is issued. The optimal policy turns out to be one in which the server retries the connection one or more times before embarking on the timeout period. This reduces the expected time to success (since retrials may take different connection paths) while assigning little additional work to the server. The effectiveness of such a policy is certainly familiar to anyone who has attempted to connect to a slow website.

Yet another application of the retrial queueing model to computer networking, and particularly to wireless networking, is described in [110]. The problem in this case is to determine a protocol that will facilitate communication between clients and servers in a *multiple-access*network, which denotes a system of networked clients that use a single communication channel.[4] Since clear transmission frequencies are hard to come by, it is often necessary to resort to the use of a single channel for transmitting data packets. As a result, there must be some form of protocol in place that permits the network to function despite the inevitability of cross-interference that occurs when two or more clients attempt to transmit through the same channel simultaneously. There are several such protocols that are being, or have been used, during the past 40 years, the first and most straightforward of which is known as the ALOHA protocol, which was developed at the University of Hawaii during the early 1970s. When two computers in a network utilizing this protocol transmit a packet

[4] The 'channel', as it is known in computer networking terminology, is synonymous with the queue in the corresponding stochastic model. In the context of wireless networking, the channel corresponds to the radio frequency used by the network for communication amongst clients.

at the same time, the subsequent interference is detected by both computers, after which each independently waits a random time before re-transmitting their packets. The inefficiency of this method results from the fact that the *whole* packet is always transmitted before detection may take place. The Ethernet protocol, on the other hand, uses what is called the 'Carrier sense multiple access with collision detection', or CSMA-CD, to sense when packets are in the transmission channel. This enables the network to quickly detect the occurrence of simultaneous transmissions and then terminate the transmissions *before* a complete packet has been sent.

No system resides or operates in a vacuum, and information systems are likewise influenced by an environment, which may be described as being either physical (hardware-related) or virtual (network-related). Of a certainty, the most critical and/or fragile systems must physically reside in controlled environments in order to protect the servers and network hardware from the untempered elements. That we take these precautions stems from our knowledge that the net effect of various factors such as temperature, humidity, and shock results in a *degradation* of the hardware over time, or may even cause sudden or catastrophic failure to occur. Furthermore, certain combinations of environmental factors inflict such damage at greater rates than others, and so we bank our investment in hardware on the assumption that controlling the environment, in other words, regulating the stochastic factors of temperature, humidity, etc., will mitigate the risk of total system failure.

In the context of networking, it is just as important to understand the *user* environment, for this determines the patterns of data flow, and thus the frequency of occurrence of certain events such as packet collisions across a channel. Larger networks tend to experience such phenomena in direct proportion to the number of users due to the fact that storage and link capacities are finite, and expensive at that. Moreover, variations in the number of users at different times will likely affect the speed at which information is disseminated within, or transferred out of, a given network. In developing a stochastic model of such a system, one may opt to ignore the minutiae of these variations by assigning average rates of data packet arrivals and transmission; indeed, this may be sufficient if the network is small, sees little variation, or has a large capacity with respect to the number of authorized users. In other scenarios, however, the greater sensitivity of system parameters to such variations makes it necessary to include them in the model. An example of such a system would be a group of web servers that serve a nationwide customer base. Such systems often operate at or near capacity during peak hours, and the resulting strain on the system exhibits itself in dramatically decreased data-transfer rates. At other times, the transfer rates will be at the maximum that the physical hardware can support. Modeling such networks with an adequate degree of fidelity may thus entail the adjustment of service rates based on the time-of-day.

One alternative for incorporating the effects of *external* environmental factors into a stochastic model is by using a *random environment.* The random environment process is itself a stochastic process whose state is usually assumed to be independent of the state of the process that it influences. It may take a number of varied forms such as a discrete- or continuous-time Markov chain, a random walk, a semi-Markov process, or Brownian motion, to name a few. These mathematical objects are also called *modulating* processes. Consequently, if the random environment is Markovian, the primary stochastic process to which it is attached is said to be *Markov-modulated.* One might also see the key phrases 'varying randomly', 'state-dependent' (if one is referring to the state of an *external* modulating process, not an internal system state), or 'in a Markovian environment'. There are other synonyms for the same exogenous environmental process.

A random environment can be used to modulate any of the system parameters, which in the context of a retrial queue queue might include the arrival, service, or retrial rates. If another process – such as one that provides the timing of server failures – feeds into the same system, then the random environment may be used to modulate its parameters as well. As may be expected, the dependence of system parameters upon the random environment is implemented via the use of functional relationships with time as the dependent variable. The most straightforward way to define such a function is to simply assign rates deterministically according to the state of the random environment. More uncertainty may be introduced into the stochastic model by constructing a process that assigns the rates randomly according to a stream taken from a probability distribution. Regardless, care must be taken to distinguish the external random process from the *internal* version in which the rates depend upon the states of the primary system. This latter object is often called a *state-dependent* model, although it appears in the literature occasionally in the random environment context.

It is easy to see the relevance of this combination of the retrial queueing model and an external random environment. For a physical environment, one may use a degradation model in which the effects of heat, humidity, radiation, and other effects produce cumulative damage, to include electronic hardware and other, more solid-state items. For the ethernet model, one may have incidents of network congestion occur on the heels of arrivals in, say, an external Poisson process, or the transmission rate may vary according to the state of a semi-Markov process. If one is confident that the random process modulates rates or otherwise behaves in a manner similar to that of the actual system, then one might expect that the state measures derived from such a model will accurately predict how the system will behave under the assumed conditions.

1.2 Research Objectives

Owing to the utility and interesting mathematical properties of retrial queueing models, a vast literature on the subject has emerged over the past several decades. However, relatively few researchers have touched upon the subject of unreliable servers in the retrial context, and even fewer have considered the impact of a randomly evolving operating environment on the performance measures of such systems.

A primary goal of this research is to advance the theory of retrial queues by considering systems whose operating parameters (e.g., arrival, service, retrial, failure and repair rates) are modulated by a time-varying environment. The main objectives of this dissertation can, therefore, be summarized as follows:

1. To develop basic results, such as stability conditions and performance measures, for unreliable, single-server retrial queues that operate in a random environment assuming customers bring an exponentially distributed service requirement to the system;
2. To extend the results obtained in Objective (1) to consider systems in which customers bring generally distributed service requirements to the system;
3. To illustrate how the main results can be used to improve the performance of such systems by optimally selecting operating parameters that minimize the steady-state mean time spent in the retrial orbit.

1.3 Dissertation Outline

In the next chapter, we review the literature pertinent to general retrial queues, retrial queues with unreliable servers, and general (non-retrial) queues operating in a random environment. In Chapter 3, the rudimentary concepts of the class of quais-birth-and-death (QBD) processes, as well as Markov chains of the $M/G/1$-type, are reviewed. These concepts serve as a foundation for the main stability result of the exponential model in Chapter 4 which provides valuable insights into the conditions needed for positive recurrence of generalized *level-dependent QBD processes*. Chapter 5 extends the results of Chapter 4 by considering customers who bring a generally distributed service requirement, an extension that adds considerable complexity to the analysis of the model. This model possesses an embedded *level-dependent Markov chain of M/G/1-type*, and we prove conditions for the stability of this complex system. The sixth chapter illustrates how the results of Chapter 4 can be used to enhance the performance of unreliable retrial queueing systems by optimizing the arrival and service rates associated to the distinct environment states. The seventh and final chapter summarizes the main contributions of this dissertation and provides some directions for future work in this area.

2. Review of the Literature

Retrial queueing comprises a significant portion of the modern literature on queueing theory. For a general survey of retrial queues and a summary of many results, the reader is directed to the works of Yang and Templeton [113], Falin [38] and references therein. The wide range of applications of retrial queues with unreliable servers has generated much interest among stochastic systems researchers. On a seemingly different note, the fundamental relationship between the reliability of an item and its working environment has popularized research into *random environments* since the late 1940s. The high level of attention that each of these separate subject areas has received has not, however, produced the simultaneous pairing of retrial queueing systems with unreliable servers whose parameters are modulated by a random environment. In this chapter, we shall review the literature that pertains to these unrelated research tracks and the emerging literature that begins to suggest their combination.

2.1 Analysis and Control of Retrial Queues

A *retrial queue*, as defined by Falin in [38], is essentially a queueing system with no waiting room in which blocked entities, or customers, may revisit the server at some random time in the future. Thus, these secondary (retrial) entities, do not leave the system, but instead proceed to a buffer, which in the context of retrial queues is termed an *orbit.* No stipulation is placed on the structure of the orbit nor the policies that specify how entities retry the server or how they are released from the system. In particular, the leeway granted in determining the nature of the blocking, whether it be a busy or failed server, and the mechanism by which blockings occurs leads to a wide variety in the open literature on the topic. In general, however, the structure of the orbit takes one of two forms. The first is a queueing structure for which entities are released back to the server according to a first-in-first-out (FIFO) discipline. The other type is that of an infinite-server queue in which each retrial entity revisits the server after some randomly-distributed duration but independently of every other retrial entity. This paradigm is appropriate for modeling such real-world systems as computer networks in which the entities – data packets – are homogeneous and no ordering protocol is defined.

The seminal papers in retrial queueing were published by Kosten [58], Clos [30], Wilkinson [112], and Cohen [31, 32], each of whom were working to solve blocking issues in telephone networks. The key to their research was the determination of performance measures related to blocking probabilities; for example, the probability of having n busy trunks (primary paths in a telephone network) and the percentage of dropped calls. In a 1957 paper by Cohen [32], the entities are calls that possess independent and identically distributed (i.i.d.) general inter-arrival times with exponential holding times (in service). Calls may take any idle trunk; if no such trunk exists, then the call may either be discarded or sent to an *overflow trunk* - which is, of course, the retrial orbit. This system is useful in modeling a telephone caller who, once blocked by the system from placing his call, may redial at some indeterminate time in the future. As Clos [30] observes, the preponderance of callers will in fact try to call again, which is synonymous with 'will not leave the system'. See [30, 31, 32, 58, 112] for further details on the authors' respective models and analytical approaches.

The retrial process as envisioned by Kosten [58], Clos [30], Wilkinson [112] was extended to single-server queues with no waiting room and general *service* times – in other words, a retrial queue based on the $M/G/1/1$ model – by Keilson, Cozzolino, and Young [48]. Retrial times, as well as primary inter-arrival times, follow exponential distributions. The authors derive the basic asymptotic measures for the one-server case based upon transient results derived by Keilson and Kooharian in [49] for the standard $M/G/1$ queue.

The next significant result was published in 1983 by V.G. Kulkarni [60]. In this article, he proves a seemingly simple equilibrium result $\lambda R = \lambda_s R_s$, which basically states that, in the long run, the mean number of total unsuccessful attempts by incoming customers equals the number of such attempts during service periods. He then uses this fact to obtain the mean number of differentiated customers in each of two classes in an $M/G/1$ system with different exponential retrial times and arrival rates. Choi and Park [29] follow on the heels of this result in their 1990 paper. Their retrial model consists of a primary queue and orbit, each of infinite capacity, and a Bernoulli splitting mechanism that routes a proportion of blocked customers to the orbit. Falin and Artalejo [39] expand the analysis of finite source retrial queues[1] to include the waiting time process and an excursion into transient measures via the busy period. Finally, Kumar and Arivudainambi [64] retrial queues consider the $M/G/1$ retrial model in which server vacations are controlled by a Bernoulli process and the orbit is governed by a FIFO discipline in which only the customer at the head of the orbit queue is allowed to access the server. The Bernoulli vacation process is of special interest as it relates directly to the topic of failures in retrial queues, which we shall introduce next.

2.2 Retrial Queues With Unreliable Servers

In the conventional retrial queueing model, customers proceed to the orbit if they find all of the servers busy upon entering the system. It is a natural extension to consider failures of the server, particularly in light of the relevance of such a model to the reliability study of real-world systems. Much of the terminology in the literature regarding ordinary queues with unreliable servers has passed on to their counterparts in retrial queues. In this case, it becomes important to distinguish between failures that occur *during* service ('active breakdowns') and those that occur while the server is idle ('nonactive breakdowns') since this will obviously impact the

[1] In other words, there exist a finite number of possible arrivals to the queue.

orbit size. Realistically speaking, failures occur probabilistically, with failure states represented in the model by states in a stochastic process (such as a Markov chain). This scheme may be simplified further via the use of *alternating states* in which a two-state stochastic process governs failure in an on-off fashion. Accordingly, the server is assigned a rate of repair that is independent of the retrial rate of customers in the orbit, while arrival and service rates depend upon the state of a discrete modulating process. This creates an interesting dynamic in which such measures as the orbit size, customer throughput, and sojourn time, become the primary focus of investigation.

The first mention of retrial queues with unreliable servers appeared in an article by Yang and Templeton [113], which is a survey of retrial queues in the spirit of Falin [38], but which additionally associates the breakdown-related topic of retrial queues to server vacations.[2] Nevertheless, Aissani [4] is credited with developing the first retrial queueing model[3] that included server breakdowns in his seminal article published in 1988. The author introduced two variations of the unreliable retrial model: (1) the queue whose server breakdowns occurs according to a Poisson process, and (2) the same queue, but with failures determined by a two-state Markov chain. He then generates the performance measures of interest via probability generating functions (p.g.f.).

Two years later, Kulkarni and Choi [62] independently derived results for an $M/G/1/1$ retrial queue with exponentially-distributed retrial and failure times. Just as in [60], the failure rate is modulated by the status of the server, where by 'status' the authors mean 'idle' or 'busy'. In the meantime, Aissani [4] continued to build upon the results he had obtained up to 1988. In the article that follows, Aissani [5] re-examined the first of the two models studied previously in [4] by generalizing the distributions of the failure times and by specifying batch Poisson arrivals. He demon-

[2] The distinction here is that *vacations* are server breakdowns that occur when the server is *idle*.

[3] The author refers to retrials here as *repeated calls*. Other synonyms for *retrial* include *repeated orders*, *repeated attempt*, and *returning customer*.

strated that the ergodicity conditions hold for the corresponding three-dimensional system evolution process. He then derived an expression for the z-transform of the long-run size of the secondary-source buffer (orbit).

Aissani [6] continued to elaborate upon his model with the addition of redundant servers that substitute for failed primary servers during a corresponding repair epoch. This was followed by a paper that he co-authored with Artalejo in 1998 [7] in which he introduces a modified version of the $M/G/1/1$ queue that appeared in its original form in both [4] and [62]. In that paper, the system is subject to exponentially-distributed server failures in which distinct failure rates apply according to whether the server is busy or idle. They also introduced the *auxiliary queueing system* in which interrupted customers have the choice either to leave the system or to remain. Lastly, the authors introduced the concept and terminology of the *fundamental server period*, which is defined as the period of time between the start of service and the next time that the server is available to begin processing another customer. Note that the interval of time corresponding to a fundamental server period does not necessarily end with the conclusion of the current customer's service in an unreliable retrial system. To be more specific, if service is interrupted, then this interval of time ends at the instant that the server becomes operational (i.e. after repair).

The years following 1994 are marked by steady progress in the development of a variety of queues with unreliable servers as evidenced by the numerous contributions published during that time period. Of notable mention is the work of Sherman and Kharoufeh [96], in which results are derived for an $M/M/1$ retrial queue with unreliable server and infinite-capacity primary queue and retrial orbit. The authors derive explicit expressions for the limiting distributions of orbit size, queue size, and number in the system via a p.g.f. approach. Other useful references from this fruitful period of time can be found in [9, 11, 14, 35, 42, 52, 59, 63, 65, 70, 71, 97, 102, 111].

2.3 Queueing Systems in a Random Environment

In the context of queueing theory, the term 'random environment' refers to a stochastic process that controls, or *modulates*, one or more of the parameters of the queue. In other words, it is an exogenous process whose evolution is independent of the process it modulates. If the external random process is Markovian, then the primary stochastic process is said to be *Markov-modulated*, and it is this form of random environment that is most often seen in the queueing literature, though it is certainly not the only one. The number of classifications into which the work to-date on stochastic systems operating in random environments is vast. Hence, the reader is referred to [21, 34, 43, 55, 56, 90, 107] for further reading.

The modulating random process possesses the natural interpretation of being an external environment that instigates change in some system parameter of interest. As an example, one might consider the effects of ambient temperature and humidity on the proper operation of a circuit board to be an interpretation of the analytical notion of an external environment. This connection makes queueing in a random environment critical to the analytical modeling of real-world environmental effects, and which leads to the alternate moniker 'random environment'. In what follows, we will discuss the important trends and results over the past 40 years since the seminal papers on random processes in queueing have appeared. This will make apparent the current open problems involving queues in random environments, and thus, pave the way for the contributions that will be made in this dissertation.

Random environments, as they appear in the literature, are cast into the form of virtually every stochastic process known using every possible modulating mechanism. The earliest methods for incorporating a random environment into a stochastic model were either to use a random draw of service (or arrival) rates discretely (using Bernoulli trials) or in accordance with some continuous distribution. Later, finite-state Markov chains, and even semi-Markov processes, were employed as modulating processes. Accordingly, random environments appeared under a variety of names, such as ‘variable service and arrivals’ and even as specialized as ‘wear processes’. Researchers have considered nearly every possible facet of these environments, not only for their intrinsic interest, but also for the fidelity that these models afford to analysts studying real-world systems.

The seminal articles on this subject were published by four sets of authors over a period of a decade. The first batch that appeared in 1963 consisted of works by Eisen and Tainiter [37], Avi-Itzhak [18], and Avi-Itzhak and Naor [19]. This was followed in 1966 by the work of Leese and Boyd [69], and, finally, in 1971 by Yechiali and Naor [115], who independently worked on the same problem as the one that appeared in [37]. After the initial foundations were laid in the intervening ten-year period since 1963, researchers began to experiment with various combinations of stochastic processes and random environments. Stochastic researchers from other fields picked up the topic, and soon thereafter, it became a staple in the stochastic elements of every physical science, and particularly all fields of engineering and reliability.

Among the first articles on stochastic systems in a random environment is that of Eisen and Taineter [37]. Their model is a single-server Markovian queue whose rates are modulated by a two-state process with negative-exponential inter-transition intervals. The equilibrium state measures were computed via the solution of a system of differential equations obtained, in part, from the Kolmogorov forward equations. Though Eisen and Tainiter are given credit by Purdue [88] as being the first to incorporate variable arrival *and* service rates into a queueing model, there were others working concurrently on related issues. The queueing model described by Avi-Itzhak [18] in the second of a two-part article is priority-based with heterogenous arrivals, but with service rates chosen at random (i.e. according to a probability distribution) at each arrival. This system also included server breakdowns, but customers avoided being pre-empted by repeatedly choosing service times until a sufficiently short service interval is obtained. This paper comes closest in spirit to the research that we present here, namely in that if preemptions were allowed, then the system becomes a retrial system in a random environment subject to breakdowns.

The interesting feature of Leese and Boyd's work [69] is that they present numerical techniques for the computation of *transient* measures of a simple $M/M/1$ queue in which the service rate depends explicitly on time. On the other hand, Yechiali and Naor [115] independently arrive at the same results for the same two-state model as in Eisen and Tainiter's paper in [37], after which Yechiali generalized this work to the case of greater than two states in [114]. Neuts [79] developed transient as well as steady-state results for a Markov-modulated $M/G/1$ system in which service rates are fixed for the duration of service. This model sets the paradigm – namely that system transitions may not occur simultaneously – that governed his later work [80] and the Markov-modulated queueing systems studied by other researchers in recent years. Finally, Purdue [88] incorporated all of the pioneering work heretofore mentioned into a rigorous analysis of the $M/M/1$ queue in a Markovian environment in an article published in 1974. In it, he derived one of the earliest results for the busy period[4] when the Markovian environment possesses greater than two states. There are also a number of other papers during this timeframe that deal with service- or interarrival-time dependency on an external process; see [44] for a look at how the random-environment mechanism evolved over its initial 20-year period.

In 1976, Kogan and Litvin [56] considered finite-capacity queues in a two-state Markov random environment and derived the associated stationary measures of mean queue length and the probability of *service failure* via transforms. Although the authors intended 'service failure' to mean a failure in throughput (i.e. probability of a full queue), this was the first time that anyone had ever considered the probability of an undesired event in random environments. In 1978, Neuts again published two articles [84, 85] concerning random environments in queueing. The first in

[4] This denotes the contiguous period of time between two epochs in which the queue is empty.

the series focused upon the same type of $M/M/1$ queue in a random environment considered in [37]. This time, however, he considered the driving Markov chain in the framework of the quasi-birth-and-death process (QBD) and then applied the matrix-analytic treatment to compute the steady-state measures of interest. This paper reflected, among other things, Neuts' predilection for methods that allow for the algorithmic computation of state measures in the elegant matrix-analytic fashion. In the second half of this research (i.e., the second article), Neuts contributed two important results. The first is that the equilibrium queue-length distribution at the end of a sojourn is the same as that any time *during* the sojourn. The second important development is that he derived steady-state results for the multi-server queue $M/M/c$, which is the first time that anyone has accomplished any analysis for a multi-server queue in a random environment.

During the latter part of the 1970s, a good degree of specialization took place in the literature as various authors began to explore a range of issues that were already known for queues with constant parameters. Many, if not the majority of these earlier works were done with a type of single-server queue since the analyses were rather involved even for the $M/M/1$ in a random environment. In 1984, Mokaddis, Elias, and Metwally [77] studied the $M/M/1$ bulk-service system whose service and arrival rates are subject to modulation by a bivariate Poisson process. A bulk-service system is one in which an idle server will take exactly $r \geq 1$ customers from the queue and serve them *en masse* according to a randomly-assigned service time. Mokaddis, Elias, and Metwally [78] likewise considered an $M/M/1$ system, but this time with an unreliable server in a random environment. They derived the mean queue length and probability of service failure using a partial-generating function technique.

Some authors extended their considerations to queues with arbitrary arrival and/or service-time distributions. Baccelli and Makowski [20] in 1986 obtained stability conditions for the $G/G/1$ queue with service subject to a random process. Their conclusion was that the workload in queue was larger for the corresponding system with convex ordering than for one with a deterministic service rate based on the average service time. Of significant notice is mention of the intensity-conservation laws for queues that were derived by Miyazawa [76] in 1985, which were employed in order to place bounds on the workload process. The authors also mentioned Rolski's article [93] in which the the stability conditions are derived for single-server queues with an *ergodically stable sequence of random variables* (see [27] for a concise definition) forming the nonstationary input process. Rolski then proved that the average waiting time is greater than that of the corresponding $M/G/1$ queue with the same arrival and service intensities.

Much work had been accomplished for modulating processes of single-server queues with exponentially-distributed input and service times. However, a number of authors did consider service or arrival processes that were a deterministic function of the states of the modulating process. In a very recent article, Mahabhashyam and Gautam [73] studied a single-server Poisson-fed queue with infinite capacity and FIFO discipline. Of significant note here are the assumptions that (i) if the environment process is a continuous-time Markov chain (CTMC) $\{Z(t) : t \geq 0\}$, then the service is performed at a rate of $b_{Z(t)}$ units of work per unit time; in other words, the work is constant between transitions of the modulating CTMC, and (ii) state transitions of the CTMC are permitted *during* service. They derived the first and second moments of the service time using first-step analysis, with the conclusion that the moments of service time are dependent upon the arrival process. In order to compute the essential measures of queue length and waiting time[5] the authors resorted to a matrix-geometric approach. They completed their discussion with applications to networks and CPU processor-sharing. The reader is referred to [26, 91, 94, 109] for further reading.

[5] The distributions of queue length and waiting time cannot be obtained in closed form for this system.

As is usually the case, the derivation of steady-state measures abounds in the literature due to complexity issues that plague the derivation of the time-dependent, or *transient* measures of a queue, and especially one that is modulated by an external process. In 1992, Lee and Li [68] considered a Markov-modulated Poisson-arrival queue under *overload control.* That is, when the buffer content exceeds a certain level, the arrival process is modified in order to mitigate any overflow condition. The authors determined the transient distribution of queue length and the first passage time to and from overload status for this model. They then determined the optimal conditions under which the buffer content rises as slowly as possible and decreases as quickly as possible. Finally, the authors studied how the properties of the modulating Markov chain affect the transient behavior with regard to the aforementioned state measures.

Markov-modulated queues using the matrix-geometric approach to queueing developed by M.F. Neuts during the 1970s has created an entirely new discipline because of the fresh approach that it offers to existing problems that are intractible under analysis by conventional means. In 2005, Mitrani [75] applied matrix-geometric techniques to unbounded queueing systems in Markovian environments. His stated goal was to obtain suitable approximations for the exact state measures of the system under a heavy load and subject to an environment with many states. As the author explains, existing methods for computing exact measures suffer from numerical instability and complexity due to such factors as ill-conditioned matrix terms. The proposed method centered around a derived matrix expression $Q(x)$ and a proposition concerning this matrix. In simple terms, the proposition states that the associated QBD for the Markov-modulated system is ergodic if and only if the number of eigenvalues of $Q(x)$ in the unit disk is equal to the number of states of the Markovian environment process. From this, one can conclude that the joint distribution of the queue size and environment is close to being geometrically distributed, with parameter identical to the dominant eigenvector of $Q(x)$. This, of course, leads naturally to a method by which one may approximately ascertain the queue length.

It is worth mentioning that the notion of *fluid queueing* was being developed in conjunction with the random environment concept. In fact, fluid queueing owes its existence to the notion of a random environment since rates of input flow of a continuous fluid are typically modulated by an external stochastic process.[6] It is this continuous input that distinguishes the fluid system from the standard queueing system. Nevertheless, it becomes apparent that the fluid queue may be (roughly) analogized to an *infinite*-server queueing system that is Markovian if its rate-modulating process is likewise Markovian. Kella and Whitt [50] uncover yet another analogy to standard queueing in considering a fluid system, and, namely one with a type of compound modulating process. The fluid queue alternates between 'up' and 'down' states; when the system is up, the buffer content decreases according to one stochastic process and when it is down, the buffer content increases according to another process. The authors then demonstrated that the steady-state buffer content is directly related to the virtual waiting time in a $G/G/1$ queue under certain assumptions.

Non-queueing systems in random environments have also been considered, and although these types of environment-modulated processes are not the focus of this research, the methods that they use to derive associated performance measures are highly relevant. In 1981, Bourgin and Cogburn [24], discussed the probability of passage into a closed set of absorbing states for a Markov chain in a random environment. Economou [36] took a much less esoteric approach in deriving the stationary distributions for measures of a bivariate discrete Markov chain $\{(E_n, X_n) : n \geq 0\}$, of which the first term is taken from the random environment $\{E_n : n \geq 0\}$. A year later, Hu [46] considered a Markov chain, a renewal process, and a random walk – as well as a queue – in a random environment and established each of their state space decompositions. In 1992, Korotaev and Spivak [57] considered finite-capacity

[6] The output rate is usually deterministic.

queues with parameters that vary according to a semi-Markov process. Moreover, the system state changes very slowly, with sojourn times in an inverse-reciprocal relationship with a small positive ϵ. The authors obtained the distribution of the number in the system via solution of coefficients in a series expansion.

Reliability analysis in queueing theory is a natural direction for the research into failure models to take. We see the first set of such works for reliability systems (mainly pioneered by Sztrik; see [10, 101, 103, 104, 105, 106]) scheduling, and financial stochastic systems and other logistic considerations appear in this context from 1989 onward; see [28, 33, 45, 47, 51, 86, 99] for specific examples.

Of immediate relevance is the topic of Markov-dependent single-server queues; that is, those queueing systems whose exponential service *and* arrival processes depend upon each other. The literature here is quite extensive, but does not necessarily contain works in which the inter-arrival and service times are *modulated*; nor do they deal with server breakdowns. One of the more recent works in this area is by Adan and Kulkarni [3], who study the limiting distributions of the waiting time and queue length of a semi-Markov queue whose inter-arrival and service times are modulated by the same discrete-time Markov chain. The assumptions are no more stringent than that of the queue possessing the Markov property, which considerably simplifies analysis. Hence, the importance of this work is that the authors consider single-server Markovian queues with general distributions and subject to a random environment.

2.4 Unreliable Queueing Systems in Random Environments

There does not exist an abundance of papers in the literature that deal with queues in random environments *and* the possibility of service interruptions and/or vacations, at least explicitly. The idea was recognized early, as Avi-Itzhak did in his 1963 articles [17, 18], but obviously presented too many challenges at such an early time in the development of the subject. Regardless, failures are implicit in any practical model of a real-world system, and so the foundations for the modulated queue with failure were established almost as soon as the first articles on models with process-dependent parameters. In 1963, Avi-Itzhak and Naor [19] describe five different single-server queues that incorporate different assumptions about the failure. The first failure model is based upon a stationary Poisson process with no restrictions. The second model assumes that failures can occur only during a busy period, the third assumes that failures occur when the system is nonempty, the fourth assumes that repair takes place only at the request of a customer, and the last assumes that failures occur only during idle periods. The repair and and service times are both generally-distributed with density and finite second-moments for each model.

While this work established a methodology for the incorporation of failure into a model, some time passed before these were found in conjunction with variable rates. In 1976, Kogan and Litvin [56] computed the asymptotic measures for a queueing system in an unspecified random environment and subject to service failures. Mokaddis, Elias, and Metwally [77, 78] did the same for modified $M/M/1$ queues with Poisson-modulated rates. In 1999, Kroese and Nicola [59] considered single-server (fluid and discrete) queues with alternating failures and Markov modulation of the Poisson arrivals and generally-distributed service rate. They employ results on Markov-additive processes to obtain results on the *optimal change of measure*, and the concept of the *effective bandwidth* is used to restrict the number of environmental states that need to be included.

Finally, in 2005, Klimenok [54] studied what is essentially the first known retrial queueing system in a random environment to appear in the open literature. The model is comprised of a single-server with a batch Markovian arrival process (BMAP) and semi-Markov service and retrial intervals. The process that defines the random environment is a bivariate Markov chain $\{(r_t, s_t) : t \geq 0\}$ with a finite state space. The random variables η_{r_t,s_t} of customers served sequentially are influenced by the random environment by definining them to be geometric with parameter $q_{r,t}$. The first term of the bivariate process controls a hybrid mechanism in which the queue is determined to be either a retrial queue or a 'system with waiting' based on its membership in a partition of the subsets $\{(r,\cdot)\}$ of the state space. The second term controls the parameters of the BMAP input and the semi-Markov intensities of service and retrial. It is a *synchronous* random environment in the sense that it only changes its state at service completions, thus obviating the need to consider changes in the service rate during transitions of the random environment. The author then employed an embedded Markov chain to evaluate the queue in steady-state, and thus, derive the associated distributions of state measures as probability generating functions.

Aside from [54], the literature concerning Markov-modulated retrial queueing systems is, at best, sparse, and such a model that includes failures of the server(s) does not exist to the author's knowledge. Thus, it is the aim of this research to supplement the retrial queueing literature with novel insights into the stability and steady-state behavior of a class of models that has not been considered previously, namely the $M/M/1$ and the $M/G/1$ versions of the unreliable retrial queueing system in a random environment. Moreover, it is crucial that the analysis presented here be useful for practical application, which suggests that our approach must be oriented to computational considerations and algorithmic development. The matrix-analytic theory turns out to be an ideal framework for this purpose. It is firmly grounded in mathematical principles, and, yet, is easily utilized in the computational investigation of queueing performance. In this dissertation, we seek to not only apply the matrix-analytic methods, but also to extend their applicability to the larger classes of level-dependent $GI/M/1$ and $M/G/1$-type systems.

3. Preliminaries

Markov chains play a fundamental role in the theory of queues, and particularly those that can be categorized as *birth-and-death processes.* These are continous-time Markov chains (CTMCs) for whom transitions are allowed only to neighboring states. Suppose that we are given a homogeneous CTMC $\{X(t) : t \geq 0\}$, where $X(t)$ denotes the population at time t. The state space of this CTMC is $S = \mathbb{Z}^+$, which are the nonnegative integers, and $Q = [q_{ij}]$ is its infinitesimal generator . When the population is i, that is $X(t) = i$, then the exponential rate of *births* is λ_i and the rate of deaths is μ_i. In mathematical terms, this translates to

$$\begin{aligned} q_{i,i+1} &= \lambda_i, \quad \text{if } i \geq 0 \\ q_{i,i-1} &= \mu_i, \quad \text{if } i \geq 1 \\ q_{ij} &= 0 \quad \text{otherwise.} \end{aligned}$$

The transition rate diagram for a birth-and-death process is shown in Figure 3.1. As a consequence of the definition of the transition rates, we may write the generator Q in the following manner:

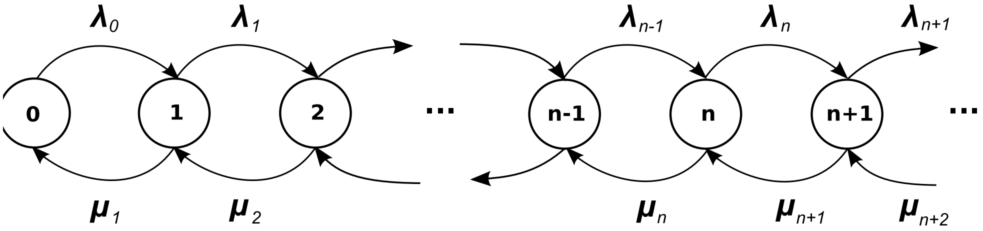

Figure 3.1 Transition rate diagram for a standard birth-and-death process.

$$Q = \begin{bmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & 0 & 0 & \cdots \\ \mu_1 & -(\lambda_1+\mu_1) & \lambda_1 & 0 & 0 & 0 & \cdots \\ 0 & \mu_2 & -(\lambda_2+\mu_2) & \lambda_2 & 0 & 0 & \cdots \\ 0 & 0 & \mu_3 & -(\lambda_3+\mu_3) & \lambda_3 & 0 & \cdots \\ 0 & 0 & 0 & \mu_4 & -(\lambda_4+\mu_4) & \lambda_4 & \cdots \\ 0 & 0 & 0 & 0 & \mu_5 & -(\lambda_5+\mu_5) & \ddots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

Let p_j be the probability that $j \in \mathbb{Z}^+$ is the population of the system at steady-state. The existence of the steady state distribution for a birth-death system hinges upon the existence of a solution to the system of equations

$$\boldsymbol{p}Q = \mathbf{0}, \qquad \boldsymbol{pe} = 1,$$

where $\lambda_i > 0$, $\mu_i > 0$, $i \in \mathbb{Z}^+$, and $\boldsymbol{e}$ is a row vector containing ones. If a solution exists, it is given by

$$p_0 = \left[\sum_{j=0}^{\infty} p_j\right]^{-1} \tag{3.1}$$

$$p_i = \pi_i p_0, \quad i \geq 1, \tag{3.2}$$

where

$$\pi_i = \prod_{j=0}^{i-1} \frac{\lambda_j}{\mu_{j+1}}, \quad i \geq 1.$$

The product form (3.2) of the steady-state probabilities, in particular, is a hallmark of all birth-and-death processes. This relationship to state 0 is of fundamental importance to these and other more generally-defined quasi-birth-and-death (QBD) processes that we shall discuss next, and is the cornerstone of the matrix-analytic approach as it pertains to these models.

Many Markovian queueing systems are modeled as birth- and-death processes. For example, the $M/M/c$ and $M/M/c/c$ queueing systems evolve as birth-and-death processes. However, the simple birth-and-death model fails to account for multiple interacting stochastic processes that might coexist in more complex systems, and thus, a more general paradigm has become prevalent in the queueing community. In this chapter, we review the rudimentary notion of the QBD and illustrate the techniques with which one may easily analyze such stochastic processes. Subsequently, we discuss the extension of this idea to non-Markovian processes that behave in a similar manner, but nevertheless exhibit non-exponential random behavior. These preliminary concepts are needed for the formal analysis of unreliable retrial queues that operate in a randomly evolving environment.

3.1 Quasi-Birth-and-Death (QBD) Processes

The study of quasi-birth-and-death processes is inextricably tied to the *matrix-analytic* approach of Neuts (see [81, 82, 83]). In contrast to the more conventional z-transform approach used to compute the limiting distribution of such processes, the matrix-analytic method takes advantage of a common structure that is shared amongst a wide variety of seemingly unrelated models. Such general applicability leads to an algorithmic approach. While this may seem to be a trade off, it soon becomes apparent that the method allows for the solution of realistic systems with significant levels of complexity.

Such tractability is a consequence of the fact that essential quantities, such as those that exist in (3.2), may be represented as *single matrix entities* regardless of the size or complexity of the model. This fact, when combined with the convenient *matrix-geometric* distribution of the steady-state probabilities, such as that which is embodied in (3.2), allows for the relative ease of the computation of the steady-state performance measures for models that may be classified as QBDs.

We shall next review basic definitions, concepts, and techniques of the theory of quasi-birth-and-death (QBD) processes, all of which is founded upon the matrix-geometric property of discrete- and continuous-time Markov chains. We shall initially present basic concepts that fall within the framework of discrete-time QBDs. This is done for two reasons. First is the fact that the properties of continuous-time QBDs are derived from those of their discrete-time embedded Markov chains, which are likewise QBDs. Second, probabilistic interpretations are more clearly understood and applied for the discrete case by virtue of the fact that these are defined by their transition probabilities. The important distinction is that which separates *level-independent* QBDs from those that are *level-dependent*, the latter type being not only more complex, but which also require a different set of methods for their steady-state analysis.

3.1.1 Level-Independent Processes

A bivariate DTMC $\{(X_n, Y_n) : n \geq 0\}$ is a QBD if its transition probability matrix appears in *block-tridiagonal* form

$$P = \begin{bmatrix} A_1 & A_0 & 0 & 0 & 0 & \cdots \\ A_2 & A_1 & A_0 & 0 & 0 & \cdots \\ 0 & A_2 & A_1 & A_0 & 0 & \cdots \\ 0 & 0 & A_2 & A_1 & A_0 & \cdots \\ 0 & 0 & 0 & A_2 & A_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{3.3}$$

where A_0, A_1, and A_2 are square matrices whose dimensions are determined by the structural properties of the system at hand. We further observe that the off-diagonal blocks A_0 and A_2 contain only positive entries as they provide information about the transition probabilities (or rates in the continuous-time model) of the process. Each row of P is termed a *level*, which is comprised of related system states called *phases*.

It is possible to transition in one step either to adjacent levels or to any other phase within the same level, hence the term 'quasi-birth-and-death'. Incidentally, the size of the QBD blocks $A_j, j = 0, 1, 2$ corresponds to the number of phases per level. For example, if the block size is $M \times M$ where M is a positive integer, then there are M^2 phases, and vice-versa.

A QBD is termed *level-independent* or *homogeneous* if it exhibits an infinitely repeating tridiagonal block structure such as the one represented in (3.3). Nevertheless, it is usually the case that the states that comprise level 0 may differ from those at any other level. We therefore refer to these states as the *boundary states* $\mathcal{S}^{(0)} \subset \mathcal{S}$, where $\mathcal{S}$ is the state space of the entire QBD. A crucial point to be made here is that the process

$$\{(X_n^{(0)}, Y_n^{(0)}) : n \geq 0\} \tag{3.4}$$

restricted to the boundary states at level 0 (otherwise known as the process at level 0 under *taboo* of all other levels of the QBD) is itself a Markov chain with invariant probability vector $\boldsymbol{\pi}_0$ and its own transition probability matrix (to be defined later). The importance of the process restricted to level 0 is derived from its role in determining the matrix-geometric distribution of the steady-state distribution of the overall QBD, as hinted in the preface to this chapter.

The discussion that follows is drawn from [66], which consequently may be referenced for proofs and other details that we do not present here. We begin with the assumption of homogeneity. In order to adequately convey what is meant by the term 'matrix geometric', we define the following. Let $\{(X_n, Y_n) : n \geq 0\}$ be a discrete-time level-independent QBD with transition probability matrix P and a finite number of phases $N > 0$. We define the steady-state joint probabilities π_{ij}, $i, j \in \mathbb{Z}^+$ to be the limiting values

$$\pi_{ij} = \lim_{n \to \infty} P\{X_n = i, \, Y_n = j\}.$$

Next, define $\boldsymbol{\pi}$ to be the row vector $[p_{ij}]_{i,j\in\mathbb{Z}^+}$ such that $\boldsymbol{\pi} = [\pi_{00}, \pi_{01}, \ldots, \pi_{10}, \pi_{11}, \ldots]$. It is a well-known fact that $\boldsymbol{\pi}$, if it exists, is the unique solution to the system of equations given by

$$\boldsymbol{\pi} P = \boldsymbol{\pi}\,, \qquad \boldsymbol{\pi}\boldsymbol{e} = 1,$$

where $\boldsymbol{e}$ is a column vector of ones.

We partition this vector according to levels of the QBD, which is to say that we set $\boldsymbol{\pi} \equiv [\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots]$, where $\boldsymbol{\pi}_i = [\pi_{i\,0}, \ldots, \pi_{iN}]$ for $i \geq 0$. The *matrix geometric property* of the steady state distribution is then expressed via the recurrence relation

$$\boldsymbol{\pi}_{i+1} = \boldsymbol{\pi}_0 R^i, \tag{3.5}$$

where R is a matrix such that, for each $i \geq 0$, R_{jk} $(0 \leq j, k \leq N)$ records the expected number of visits to state $(i+1, k)$ before returning to level i, given that the process starts in (i, j). The matrix R is often referred to as the *rate matrix* since it may be interpreted as the 'rate of visit' to level $i+1$.

It is apparent from (3.5) that determining $\boldsymbol{\pi}_0$ and R is key to the computation of the steady state probabilities for a given discrete-time QBD. It can be shown that the restricted process (3.4) of the QBD is a positive recurrent Markov chain if, and only if, the overall QBD process is likewise positive recurrent (see [66: Thm 5.3.1]). The restricted process (3.4) possesses the transition probability matrix $B + A_0G$ and the corresponding steady state vector $\boldsymbol{\pi}_0$. Consequently, the invariant probability vector $\boldsymbol{\pi}_0$ may be determined in the usual way as the unique positive solution to a linear system of equations as described in Theorem 3.1.

Theorem 3.1. *The vector $\boldsymbol{\pi}_0$ is the unique steady state solution to the system given by*

$$\boldsymbol{\pi}_0(B + A_0G) = \boldsymbol{\pi}_0, \tag{3.6}$$

$$\boldsymbol{\pi}_0(I - R)^{-1} = 1. \tag{3.7}$$

Notice that solving the system (3.6) – (3.7) requires explicit determination of the rate matrix R. One might attempt to solve the following matrix-quadratic equation to obtain R, although this is not practical due to the computational burden.

Theorem 3.2. *The rate matrix R is the minimal (positive) and unique solution to the matrix quadratic equation*

$$A_0 + RA_1 + R^2 A_2.$$

It is clear that this criterion may fail to be of practical use for all QBDs, save those that possess the smallest of block matrix orders. Hence, we must resort to an algorithmic approach in order to obtain the rate matrix. To this end, we will now define two other fundamental matrices.

The matrices U and G that we shall now define are important to the characterization of the steady-state distribution, and hence, to the development of algorithms to find the rate matrix R. The (j, j')-th element of the matrix G is given by

$$G_{jj'} \equiv P(\tau < \infty,\, X_\tau \;=\; (i, j') \,|\, X_0 \;=\; (i+1, j)\,), \tag{3.8}$$

where τ is a discrete random variable defined as the first passage time from level $i+1$ to level i. It may also be determined exactly as the solution to the matrix quadratic equation

$$A_2 + A_1 G + A_0 G^2 \;=\; \mathbf{0}, \tag{3.9}$$

although, just as with (3.2), computational issues make this approach impractical. The other matrix, namely $U \equiv A_1 + A_0 G$, may be interpreted as being the transition matrix of the QBD process restricted to level i until the first visit to $i-1$. The matrices R, U, and G can each be deduced from any of the others via the following mathematical relationships:

$$
\begin{aligned}
U &= A_1 + RA_2, && (3.10) \\
R &= A_0(-U)^{-1}, && (3.11) \\
G &= (-U)^{-1}A_2. && (3.12)
\end{aligned}
$$

Though rarely used to explicitly prove the positive recurrence of a QBD, Theorem 3.3 is of significant import to theoretical as well as computational considerations.

Theorem 3.3. *The QBD process given by* $\{(X_n, Y_n) : n \geq 0\}$ *is positive recurrent if and only if the matrix* G *is stochastic.*

Proof. If the matrix G were substochastic, then it must be true that, for some $j \leq N$, we must have

$$\sum_{i=1}^{N} G_{ij} = \sum_{i=1}^{N} P(\tau < \infty,\, X_\tau = (l-1, j) | X_0 = (l, i)) < 1,$$

which shows that, at best, the QBD is null recurrent. On the other hand, positive recurrence requires that all such terms must equal unity, by definition of the matrix G. □

The stochasticity of G can often be used as a convenient stopping criterion for numerical algorithms used to approximate the steady-state distribution of a QBD. One may then determine the rate matrix R through the employment of equations (3.10–3.12).

3.1.2 Level-Dependent Processes

A natural generalization of the discrete-time QBD model is to allow the blocks of P to vary according to level. A QBD with such a characteristic is called *level-*

dependent or *nonhomogeneous*, with a transition probability matrix represented by

$$P = \begin{bmatrix} A_1^{(0)} & A_0^{(0)} & 0 & 0 & 0 & \cdots \\ A_2^{(1)} & A_1^{(1)} & A_0^{(1)} & 0 & 0 & \cdots \\ 0 & A_2^{(2)} & A_1^{(2)} & A_0^{(2)} & 0 & \cdots \\ 0 & 0 & A_2^{(3)} & A_1^{(3)} & A_0^{(3)} & \cdots \\ 0 & 0 & 0 & A_2^{(4)} & A_1^{(4)} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \tag{3.13}$$

We may further relax the requirement that the off-diagonal blocks be square matrices, or even that there must exist only a finite number of phases per level. Examples of systems with level-dependent QBD representations abound in the literature. The standard Markovian single-server retrial queue with independent retrials and no buffer, for instance, has a level-dependent QBD representation that comes about due to the dependence of the total retrial rate upon the level, or number of entities attempting to reaccess the server.

It so happens that many basic facts concerning the steady-state distribution of level-independent QBDs may be extended to the level-dependent case (see [66: p 262]). For other quantities that have resisted translation from the level-independent case, such as a simple drift criterion, we may still resort to limiting processes taken over levels of the QBD. Such methods are straightforward enough in practice, but have proven to be quite difficult to make rigorous.

The classical definition of QBDs has recently evolved to include Markov chains $\{(X_n^1, X_n^2, \ldots, X_n^M) : n \geq 0\}$ with three or more state variables whose transition probability matrices assume the tridiagonal form. Choose one of the component random variables with an infinite state space, say X_n^1, and associate each of its possible states with levels of the QBD. This transforms the above multivariate process into the group-bivariate process,

$$\{\,(X_n^1, (X_n^2, \ldots, X_n^M)\,) : n \geq 0\},$$

which in turn allows the transition matrix to be represented in the usual manner by a two-dimensional matrix in tridiagonal form. In this research, we will focus upon what is called in [100] a *tri-layered QBD*, or a QBD representation of a stochastic system with three state variables. The topic of multi-layered QBDs has only recently begun to receive much attention, as the original definition of the QBD encompassed tridiagonal-structured Markov chains of only two state variables, the first of which is infinite-dimensional (the 'level') and the other, finite (the 'phase'). It became clear that the traditional QBD did not adequately fulfill every stochastic modeling task, and, moreover, that limiting the scope of QBD processes to only two-dimensions is rather restrictive.

An adequate understanding of the structure of a multi-layered QBD is crucial to the task of evaluating the steady-state performance of the system under consideration. Each so-called layer of the tri-layered QBD holds a particular significance to its analysis: the lowest (scalar-entry) level, which we shall denote here by *level 2*, is used in the numerical computation of various system measures. The middle layer, or *level 1* turns out to be critical to the determination of a closed-form stability expression. The topmost layer, *level 0*, should exhibit the tridiagonal form if the system is, indeed, a QBD.

Regardless of the number of layers involved, the matrix-geometric property for (discrete-time) level-dependent QBDs is similar to that for level-independent QBDs, with the exception that there is now a rate matrix $R^{(i)}$ that corresponds to each level $i \geq 1$. The entries of the matrix $R^{(i)}$ give the expected number of visits to level i between successive visits to level $i-1$. The vector $\boldsymbol{\pi}$ of equilibrium probabilities in this case satisfies the recurrence relation

$$\boldsymbol{\pi}_{i+1} = \boldsymbol{\pi}_i R^{(i+1)}, \; i \geq 0. \tag{3.14}$$

Accordingly, each level $i \geq 0$ is associated to distinct $U^{(i)}$ and $G^{(i)}$ matrices that obey relationships similar to those given in (3.10)– (3.12). The level-dependent counterparts to the matrix quadratic equations (3.2) and (3.9) are

$$\begin{aligned} R^{(i+1)} &= A_0^{(i)} + R^{(i+1)} A_1^{(i+1)} + R^{(i+1)} R^{(i+2)} A_2^{(i+2)}, & (3.15)\\ G^{(i+1)} &= A_2^{(i+1)} + A_1^{(i+1)} G^{(i+1)} + A_0^{(i+1)} G^{(i+2)} G^{(i+1)}. & (3.16)\end{aligned}$$

As with level-independent QBDs, the positive-recurrence of a level-dependent QBD is contingent upon the fulfillment of certain conditions by the process restricted to level 0. Theorem 3.4, which appears in [25, 66], details these requirements, as well as the subsequent matrix-geometric relationship between the resulting steady-state probabilities.

Theorem 3.4. *A continuous-time level-dependent QBD is positive recurrent if and only if there exists a positive solution to the system of equations*

$$\begin{aligned} \boldsymbol{\pi}_0 (A_1^{(0)} + A_0^{(0)} G^{(1)}) &= \mathbf{0}, & (3.17)\\ \boldsymbol{\pi}_0 \sum_{i \geq 0} \prod_{1 \leq k \leq i} R_k \boldsymbol{e} &= 1 & (3.18)\end{aligned}$$

In this case, the steady-state probability vector $\boldsymbol{\pi} = [\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots]$ *is given by*

$$\boldsymbol{\pi}_i = \boldsymbol{\pi}_0 \prod_{k=0}^{i-1} R_k, \quad i \geq 0. \tag{3.19}$$

Thus, we see that the fundamental results for level-dependent QBDs have close analogues to those already derived for the level-independent case. Nevertheless, the non-homogeneity of the QBD over its levels contributes significant complexity to algorithms devoted to the task of determining the equilibrium distribution of the QBD. We shall review these algorithms in greater detail in Chapter 4.

3.1.3 Ergodicity of QBDs

In this section we review the conditions that guarantee ergodicity of a QBD, beginning with a discussion of the requirements for a general DTMC over a countable state space. The necessary and sufficient condition for the ergodicity of a DTMC $\{X_n : n \geq 0\}$ with countable state space $\mathcal{S}$ is that the process must be *aperiodic* and *positive recurrent.* We further assume that the QBD is *irreducible* or, equivalently, that each state is reachable from all others via a finite number of transitions, a characteristic that greatly simplifies the theoretical considerations that attend the steady-state analysis.

Positive recurrence is a *class* property, which means that we shall define positive recurrence for a state $i \in \mathcal{S}$. Let S_1 be the time of the first transition of the QBD process and let $\tau_i = \min\{n \geq S_1 \,|\, X_n = i\}$ be the time of first passage into the state i. The state i is then called *positive recurrent* if the following conditions hold:

$$\text{(i)} \qquad P(\tau_i < \infty \,|\, X_0 \,=\, i) \,=\, 1, \tag{3.20}$$

$$\text{(ii)} \qquad E[\,\tau_i \,|\, X_0 \,=\, i\,] \,<\, \infty. \tag{3.21}$$

A communicating class of a Markov chain is called *positive recurrent* if every member of that class is so designated; thus, for an irreducible, positive recurrent Markov chain, *every* state must be positive recurrent. The definition of positive recurrence for a CTMC is analogous to that of the DTMC, with ‘max’ replaced by ‘inf’ in the definition of the first passage time τ_i. It is worthwhile to note here that positive recurrence is not necessarily shared between a CTMC and its embedded Markov chain *unless* it possesses a finite number of phases per level. Since we will analyze such a finite-phase model, one may assume henceforth that the positive recurrence of one implies the other.

A main focus of this research is to determine the long-run performance measures of a system that is *ergodic.* For the finite-phase, continuous-time QBD, the proof of its ergodicity is equivalent to proving that its embedded chain is likewise ergodic, which, in turn, is equivalent to proving that the embedded chain is positive recurrent (since we assume aperiodicity). Note that the determination of ergodicity is *not* contingent on the proof of the positive recurrence of every state of the embedded DTMC. This is instead a *consequence* of the *ergodicity property* for an irreducible and aperiodic Markov chain, which may be characterized as the geometric convergence of the transient probabilities of the transition probability matrix (TPM) to the steady-state probabilities; in other words, if P is the TPM and there exists an invariant probability vector $\boldsymbol{\pi}$ of the chain, and, given the column vector $\boldsymbol{e}$ of ones,

$$P^n \to \boldsymbol{e}\boldsymbol{\pi}$$

converges at a geometric rate, then the Markov chain is deemed ergodic. This fact is presented as Theorem 1.1 of [74].

For Markov chains with a countable state space, the *drift at state i* is defined as the quantity $E[X_{n+1} - X_n \,|\, X_n = i]$, where $i, n \in \mathbb{Z}^+$. A *positive* drift at i implies that the process will tend towards higher-numbered states when in state i, while a *negative* drift will imply exactly the opposite. The fundamental results that leverage this concept of drift in order to assert the ergodicity of a Markov chain are *Foster's criterion for stability* [61] and *Pakes' Lemma* [87]. Pakes' lemma, in particular, is well-known by virtue of its being the first simple criterion for determining if a discrete-time Markov chain with a countable state space is positive recurrent. Simply stated, it asserts the positive recurrence of a system that possesses negative drift over a subset of its states. Let $\mathcal{S}$ be the countable state space of a DTMC denoted by

$\{X_n : n \geq 0\}$, and let $\nu : \mathcal{S} \to [0, \infty)$. A related quantity called the *generalized drift at state i* of a DTMC is an average of weighted increments $\nu(X_{n+1}) - \nu(X_n),\ n \geq 0$, where $\nu(\cdot)$ is a nonnegative, real-valued function on $\mathbb{Z}^+$. It is formally defined for a given state i as

$$\begin{aligned} d(i) &\equiv E[\nu(X_{n+1}) - \nu(X_n) \,|\, X_n \,=\, i] \\ &= \sum_{j \in S} p_{ij}\nu(j) - \nu(i), \quad i \in \mathcal{S}. \end{aligned} \tag{3.22}$$

The function $\nu(i)$ is commonly termed a *Lyapunov* or *potential function* (see [41]), which is a measure of some positive quantity, or 'potential', associated to a particular state of a process with a countable state space. Thus, the drift may be equated to the notion of a change in potential. Foster's criterion, which we shall now state, establishes the intuitive fact that a negative net potential means that the system will not traverse its state space in an unbounded manner.

Theorem 3.5. (Foster's Criterion) *Let $\{X_n : n \geq 0\}$ be an irreducible DTMC on a countable state space S. If there exists a function $\nu : S \to [0, \infty)$, an $\varepsilon > 0$, and a finite set $H \subset S$ such that*

$$|d(i)| < \infty \qquad \text{for } i \in H, \tag{3.23}$$

$$d(i) < -\varepsilon \qquad \text{for } i \notin H, \tag{3.24}$$

then $\{X_n : n \geq 0\}$ is positive recurrent.

The form of $d(i)$ that is most relevant to our discussion of ergodicity of QBDs is $\nu(i) = i$. Indeed, Pakes' lemma, which we shall next present, may be considered as a special case of Foster's criterion in which the Lyupanov potential is the identity function.

Theorem 3.6. (Pakes' Lemma). *Let the DTMC* $\{X_n : n \geq 0\}$ *defined on the state space* $\mathcal{S} = \{0, 1, 2, \ldots\}$ *be irreducible, aperiodic, and homogeneous. Let*

$$d(i) = E[X_{n+1} - X_n \,|\, X_n = i], \quad i \in \mathcal{S}$$

and suppose that the following conditions hold for each $i \in \mathcal{S}$*:*

(i). $d(i) < \infty$,

(ii). $\limsup_{i \to \infty} d(i) < 0$.

Then $\{X_n : n \geq 0\}$ *is ergodic.*

In the next chapter we employ Pakes' lemma to describe the sufficient conditions for the stability of level-independent QBDs, where the drift is the measure of the average increment of a random walk over the state space of a QBD.

Pake's lemma, as well as Foster's criterion, is limited in that it is only an expression of the *sufficiency* of a negative drift. In [95] there appear a number of results that provide sufficient conditions for the non-ergodicity of a discrete-time Markov chain, which means the contrapositive argument may be used to obtain the necessary conditions for its ergodicity. Define p_{ij} as the (i, j)th component of the transition probability matrix P of a generic DTMC $\{X_n : n \geq 0\}$ with a state space equivalent to the nonnegative integers. It can be shown that the non-ergodicity of such a Markov chain is predicated upon demonstrating that the function

$$\psi_i(z) = \left(z^i - \sum_{j \geq 1} p_{ij} z^j \right) / (1 - z), \quad i \geq 0 \text{ and } z \in [0, 1), \tag{3.25}$$

is bounded from below by zero in some interval $[c, 1) \subseteq [0, 1)$ and for some $0 < N \leq i$. The requirement on the value of z is guaranteed by the fulfillment of *Kaplan's Condition*, which is that there exists an $\varepsilon \geq 0$, a positive integer N, and $c \in [0, 1)$ such that $\psi_i(z) \geq -\varepsilon$ for $i \geq N$ and $z \in [c, 1)$. Furthermore, the connection of $\psi_i(z)$

to drift is evident in the observation that

$$d(i) = \lim_{z \to 1^-} \psi_i(z).$$

The preceding observations culminate in the result that is Theorem 3.7, which first appeared in [95: Thm 1]:

Theorem 3.7. *If the discrete-time Markov chain $\{X_n : n \geq 0\}$ satisfies Kaplan's Condition, $d(i) < \infty$ for all $i \geq 0$, and there exists a positive number N such that $d(i) \geq 0$ for all $i \geq N$, then $\{X_n : n \geq 0\}$ is not ergodic.*

A direct proof of the applicability of Kaplan's condition may be difficult, and it is thus desirable to have alternative criteria on hand. The following theorem from [95] provides the contextual framework in which the condition holds:

Theorem 3.8. *Let $P = [p_{ij}]$ be the transition probability matrix of a DTMC $\{Y_n : n \geq 0\}$ and define the sequences*

$$\begin{aligned} \delta_i &= \sum_{j \leq i} p_{ij}(j - i) \\ \varepsilon_i &= \sum_{j > i} p_{ij}(j - i) \\ \gamma_i &= \delta_i + \varepsilon_i. \end{aligned}$$

Next, consider the following statements:

1. *There exists a k such that, for $j < i - k$ and $i > 0$, $p_{ij} = 0$.*
2. *The sequence $\{\delta_i : i = 0, 1, 2, \ldots\}$, is bounded from below.*
3. *Kaplan's condition holds.*
4. *The sequence $\{\gamma_i : i = 0, 1, 2, \ldots\}$ is bounded from below and $\lim_{i \to \infty} p_{ij} = 0$, where $j \geq 0$.*

Then $(1) \Rightarrow (2) \Rightarrow (3) \Rightarrow (4)$.

Condition (2), in particular, is well-suited to the task of demonstrating Kaplan's condition for processes, such as those of $M/G/1$-type, that allow any number of positive increments (e.g., in system size), but only a limited number of decrements.

We may now proceed to discuss the conditions that establish the ergodicity of level-independent QBDs. We note here that for the level-independent case, it is sufficient to characterize the ergodicity of a QBD in terms of its transitions between *levels*, since the phase space is identical for any given level. Thus, if one can prove the existence of an embedded Markov chain whose state space consists of the integers (and, therefore, the levels of the QBD) and whose behavior over levels is exactly that of the underlying QBD, we may then apply the preceding theorems for the ergodicity of a Markov chain.

The process that records transitions between levels bears a strong resemblance to a one-dimensional *random walk* over the nonnegative integers with discrete (independent) steps, or *increments* taken in the positive or negative directions. We shall briefly discuss the development of this idea for a discrete-time, level-independent QBD, based upon the presentation in [66: Thm 5.3.1]. Suppose that such a QBD has the transition probability matrix (3.3). We begin by defining the integer-valued random variable

$$L_n \equiv L_0 + \sum_{k=0}^{n} \delta_n, \tag{3.26}$$

for which we assume that $L_0 \geq 0$. The *increments* of the random walk process $\{L_n : n \geq 0\}$ are denoted by $\delta_n \in \{-1, 0, 1\}$ for transition epochs $n \in \mathbb{Z}^+$. Denote the phase process of the QBD to be $\{\phi_n : n \geq 0\}$ and consider the resulting Markov chains

$$\begin{aligned} W_L &= \{(L_n, \phi_n) : n \geq 0\}, \\ W_\delta &= \{(\delta_n, \phi_n) : n \geq 0\}. \end{aligned}$$

The transition probabilities of W_δ are given by

$$\begin{aligned} P(\delta_{n+1} = -1, \phi_{n+1} \ = j \,|\, \delta_n, \phi_n \ = i\,) &= (A_2)_{i,j} && (3.27)\\ P(\delta_{n+1} = 0, \phi_{n+1} \ = j \,|\, \delta_n, \phi_n \ = i\,) &= (A_1)_{i,j} && (3.28)\\ P(\delta_{n+1} = 1, \phi_{n+1} \ = j \,|\, \delta_n, \phi_n \ = i\,) &= (A_0)_{i,j}. && (3.29) \end{aligned}$$

We note here that the increments of the random walk W_L are dependent, as they arise from the transitions of the underlying QBD, and thus indicate that the process embedded at level jumps is not a Markov chain. We will deal with the dependency issue later in this discussion; our first task, however, is to determine the steady-state behavior of the increment process $\{\delta_n\}$. Through aggregation of probabilities, it is clear that the matrix $A \equiv A_0 + A_1 + A_2$ contains the transition probabilities for the phase process $\{\phi_n\}$. Thus, by choosing as the initial probability vector for $\{\phi_n\}$ the invariant probability vector $\boldsymbol{\alpha}$ corresponding to A, we initialize the process in steady-state, and thus, through association to the Markov chain W_δ, we may assume that the increment process $\{\delta_n\}$ is likewise in steady-state.

Define δ to be the limiting random increment $\delta = \lim_{n\to\infty} \delta_n$. The process W_δ, therefore, has the steady-state probability distribution vector $[\boldsymbol{\alpha}A_2, \boldsymbol{\alpha}A_1, \boldsymbol{\alpha}A_0]$. The drift

$$d(i) \equiv \lim_{n\to\infty} E[\delta_n \,|\, \delta_{n-1} = i]$$

associated to the random walk W_L for all levels is thus given by

$$\begin{aligned} d(i) &= (-1)\boldsymbol{\alpha}A_2\boldsymbol{e} + 0\cdot\boldsymbol{\alpha}A_1\boldsymbol{e} + 1\cdot\boldsymbol{\alpha}A_0\boldsymbol{e} && (3.30)\\ &= \boldsymbol{\alpha}A_0\boldsymbol{e} - \boldsymbol{\alpha}A_2\boldsymbol{e}. \end{aligned}$$

Since the QBD is homogeneous in its levels, its drift terms $d(i)$ are independent of level i, and so, for the purpose of clarity, we define $\bar{d} = d(i),\ i \geq 1$.

As previously implied, the dependence of the increments δ_n precludes $\{L_n\}$ from being a Markov chain. It is therefore necessary to extract a suitable Markovian subprocess of $\{L_n\}$ by embedding at the appropriate time epochs. As stated in the proof of [66: Thm 7.2.3], these epochs may be chosen as the times $k_0, k_1, k_2, \ldots$ of return of the process W_δ to its initial state (i.e., the first increment of the random walk $\{L_n\}$) δ_0. These instants are renewal epochs, a fact which allows us to declare that the integer-valued process $\{\mathcal{R}_n : n \geq 0\}$ embedded at such instants is a conventional random walk with independent increments. Moreover, its behavior is identical to that of the underlying QBD process until its first return to level 0, after which $\{\mathcal{R}_n\}$ may or may not decrement to a negative level. We further observe that the increments of $\mathcal{R}_n$ are given by

$$\varepsilon_n = \sum_{j=k_{n-1}}^{k_n} \delta_j,$$

a fact that allows us to compute the drift at level i as follows:

$$d_\ell(i) = E[\varepsilon_n \,|\, L_{k_{n-1}} = i] = (k_{n-1} - k_n)\bar{d}, \tag{3.31}$$

the last equality being a consequence of the fact that the QBD is homogeneous in levels. It is thus clear that $\bar{d} = d(i) < 0$ if and only if $d_\ell(i) < 0$. Hence, by Pakes' lemma, it becomes apparent that $\bar{d} < 0$ is sufficient to guarantee a finite expected sojourn time from level i back to level $i-1$ for each level i. Indeed, as is proven in [66: Thm 7.2.4], the invariance of drift over levels of the QBD also makes this a necessary condition. This enables us to state the following result:

Theorem 3.9. *Suppose that we are given an irreducible, discrete-time level-independent QBD with a finite number of phases. The QBD is positive recurrent if and only if*

$$\boldsymbol{\alpha} A_0 \boldsymbol{e} \;<\; \boldsymbol{\alpha} A_2 \boldsymbol{e}, \tag{3.32}$$

where $\boldsymbol{\alpha}$ is the unique solution to $\boldsymbol{\alpha} A \;=\; 0$, $\boldsymbol{\alpha}\boldsymbol{e} \;=\; 1$, and $A \equiv A_0 + A_1 + A_2$.

We now give a formal presentation of the *fundamental period* as described in [83: Sec 3.3], which has significant import in any discussion of the positive recurrence of a QBD and, as we shall later see, a process of $M/G/1$-type. The concept of the fundamental period is motivated by the notion of the *busy period* of a queue in the classical sense of the phrase. The busy period is loosely defined as the first passage time τ_b from the instant of the first customer arrival to the queueing system until the first instant that the system again becomes empty, which, in terms of a QBD, may also be interpreted as the first passage time from level 1 back to level 0. For the $M/M/1$ or $M/G/1$ queues, the busy period corresponds to a contiguous interval during which the server is processing work. However, this interpretation may not be valid in more complex models, particularly those in which service may be interrupted by failures or vacations. Hence, we generalize the original concept of a busy period to encompass, not just the first-passage time just mentioned, but *any* period that begins with the arrival of $i + 1$ customers to the system and ends with the system size decrementing to i.

Using the notation of [82: Sec 2.2], we begin its formal definition by introducing the first-passage times $T(i + r, j; i, j')$ from a level $i + r$, $r \geq 1$ and $Z_{i+r} = j$ to level i, $i \geq 0$ and $Z_i = j'$, where j' is the first *phase* (*hitting state*) that the process attains when it reaches level i. We likewise define the quantities $V(i + r, j; i, j')$ of the number of transitions that comprise a first-passage time $T(i + r, j; j')$, with the corresponding value for a fundamental period thus given by $V(i + 1, j; j')$. We then denote the fundamental period by the stochastic time interval $T(i + 1, j; i, j')$.

With these definitions in hand, we may finally introduce the following matrix of probabilities

$$[\hat{G}_r^{(i)}(\nu;x)]_{jj'\in S} \equiv P\left\{T(i+r,j;j') \leq x, V(i+r,j;j') = \nu\right\} \tag{3.33}$$

and its associated joint transform matrix

$$\widetilde{G}_r^{(i)}(s,z) \;=\; \sum_{\nu=0}^{\infty} z^{\nu} \int_0^{\infty} e^{-sx} G_r^{(i)}(\nu,x)\,dx. \tag{3.34}$$

Let $\hat{G}^{(i)}(\nu;x) = \hat{G}_1^{(i)}(\nu;x)$; the matrix

$$\hat{G}^{(i)}(x) \equiv \sum_{\nu=0}^{\infty} \hat{G}^{(i)}(\nu;x)$$

thus defined is called the *matrix distribution of the fundamental period at level i*. From the definitions given above, one is able to infer the relationship of the busy period to the fundamental period, which is that the matrix distribution of the busy period is given by $\hat{G}^{(0)}(\nu,x)$. We also indicate the following equivalence (the matrix $G^{(i)}$ is the level-dependent version of the matrix G defined in (3.8)):

$$G^{(i)} \;=\; \widetilde{G}^{(i)}(0,1), \quad i \in \mathbb{Z}^+.$$

Combining this observation with Theorem 3.3 formally establishes the connection between the fundamental period and the positive recurrence of QBDs.

If the QBD is level-independent, then the homogeneity of phases over all levels guarantees that $G^{(i)}(\nu,x) = G^{(j)}(\nu,x)$ for any $i,j \in \mathbb{Z}^+$. However, this is clearly not the case for a level-dependent QBD, and so we must therefore differentiate between the matrices $G^{(i)}(\nu,x)$ based upon the starting level i. The conclusion that one may draw from this discussion is that the distribution of the fundamental period varies with the starting level for level-dependent $M/G/1$-type processes. Nevertheless, the fundamental period is a *class* property, even in the level-independent case, and thus the conclusions drawn in [82: Sec 3.3] for the busy period likewise hold for the level-dependent case.

Having asserted the relevance of the level-independent conclusions, the following theorem provides a means of computing the distribution of a fundamental period for a QBD, which is a restatement in level-dependent terms of Lemma 3.3.2 and Theorem 3.3.1 in [82].

Theorem 3.10. *The transform matrices $\widetilde{G}^{(i)}(z,s)$, $z \in [0,1]$, $s \geq 0$, are the minimal nonnegative solutions to the matrix-quadratic equations*

$$X(z,s) = zC_0^{(i)}(s) + C_2^{(i)}(s)X^2(z,s), \quad i \in \mathbb{Z}^+, \tag{3.35}$$

where

$$C_0^{(i)}(s) = (sI - A_1)^{-1}A_2^{(i)} \;\; \textit{and} \;\; C_2^{(i)}(s) = (sI - A_1)^{-1}A_0^{(i)}.$$

Proof. In the proof of [83: Thm 3.3.1], the sequence $\left\{\widetilde{G}_n^{(i)}(z,s)\right\}$, $n \in \mathbb{Z}^+$, is constructed by letting $X_0(z,s) = \mathbf{0}$ in the recursive formula

$$X_{n+1}(z,s) = zC_0^{(i)}(s) + C_2^{(i)}(s)X_n^2(z,s).$$

It is subsequently proved that this nondecreasing sequence converges to the limit given by

$$\lim_{n\to\infty} \widetilde{G}_n^{(i)}(z,s) \;=\; \widetilde{G}^{(i)}(z,s).$$

□

One may thus construct a simple algorithm based on the following procedure:

$$\begin{aligned} X_0(z,s) &= \mathbf{0}, \\ X_{n+1}(z,s) &= zC_0^{(i)}(s) + C_2^{(i)}(s)X_n^2(z,s), \qquad n \geq 1. \end{aligned} \tag{3.36}$$

It is also possible to obtain numerical approximations of the matrices $\{G^{(i)},\ i \geq 0\}$ via algorithms for the computation of the steady-state distribution of a (level-dependent) QBD. Such algorithms usually produce the rate matrices $R^{(i)}$ (if not the $G^{(i)}$ matrices themselves) as by-products of the procedure. One may then employ the level-dependent version of the relationships (3.10) – (3.12) in order to obtain $G^{(i)}$. We shall discuss such algorithms in detail in Chapter 4.

3.2 $M/G/1$-Type Processes

In the previous sections, we observed how the special structure of the QBD facilitates the study of a large class of queues using a relatively modest collection of methods. It stands to reason that a matrix-analytic analog to the QBD must surely exist for classes of non-Markovian queues with highly-structured embedded Markov chains as well. Indeed, such a class, epitomized by the classical $M/G/1$ queue, has been defined, and a separate, but related set of analytical tools to those that apply to QBDs have been discovered (see [82]). In this section, we review some of the currently known solution techniques for the class of models known as *$M/G/1$-type processes*, both in terms of computing a stability criterion and also in the numeric computation of the steady-state distribution. Just as in previous sections of this chapter, we draw a distinction between level-independent and level-dependent versions of this class of models. A specific method discovered in this research for determining the ergodicity of a level-dependent $M/G/1$-type process shall be presented in Chapter 5, which addresses the $M/G/1$ retrial queueing system modulated by a random environment.

3.2.1 Markov Renewal Sequences and the $M/G/1$ Queue

The $M/G/1$ queueing system, as well as related systems of $M/G/1$-type, cannot, by virtue of its non-exponential service distribution, be described as a CTMC, nor does it possess an embedded DTMC at jump transitions in the manner of its Markovian counterpart. In order to determine the appropriate embedded DTMC (i.e. to facilitate analysis of the non-Markovian system), we resort to the theory of *Markov renewal-processes.* The idea is to construct a random bivariate sequence $\{(Y_n, S_n) : n \geq 0\}$, or *Markov renewal sequence* (MRS), at suitable epochs S_n in the system-size process $\{X(t) : t \geq 0\}$ of the $M/G/1$ queue in such a way that $\{Y_n : n \geq 0\}$ is a Markov chain, where $Y_n \equiv X(S_n^+)$. We may then study the steady-state distribution of the embedded Markov chain $\{Y_n\}$ in order to study the recurrence properties of the continous-time system.

We will not give the full details of the definition of a MRS here; for this, we refer the reader to [61: p 479]. The key characteristic is that the successive transitions of $\{(Y_n, S_n)\}$ should be independent of its history, or, in other words, the Markov property should hold for the MRS. We next let $\tau_n = S_{n+1} - S_n$ and define the probabilities

$$Q^*_{ij}(x) \;=\; P\{Y_{n+1} = j, \tau_n \leq x \,|\, Y_n = i\} \quad i,\, j \in \mathcal{S}.$$

The matrix $Q^*(x) = [Q^*_{ij}(x)]$ is called the *kernel* of the MRS. The discrete random variable $N(t)$, which is the number of observations of the MRS up to time t, defines a *Markov renewal process* $\{N(t) : t \geq 0\}$. If $N(t)$ records the number of transitions up to time t, then the process

$$\{X(t) \equiv Y_{N(t)} : t \geq 0\}$$

forms a *semi-Markov process* (SMP).

It should be emphasized that the sequence $\{(Y_n, \tau_n) : n \geq 0\}$ is also often referred to as the MRS; nevertheless, the two definitions clearly uphold the same idea with regard to the semi-Markov process that they commonly define. We now focus on the utility of the MRS, which is embodied in the statement of the following theorem which appears in [61: Thm 9.1] and is significant to the determination of the stability condition for the non-Markovian system:

Theorem 3.11. *If $\{(Y_n, S_n) : n \geq 0\}$ is a MRS, then the process $\{Y_n : n \geq 0\}$ is a DTMC with transition probability matrix given by $Q^*(\infty)$.*

We therefore wish to obtain the matrix $P \equiv Q^*(\infty)$ and explicitly define those conditions that guarantee the ergodicity of the embedded Markov chain, and thus of the original queueing system.

Just as in Neuts [82], we use the standard $M/G/1$ queue with arrival rate λ and with service distribution $H(\cdot)$ with average $1/\mu$ as a simple example of a queueing system whose embedded semi-Markov kernel demonstrates the aforementioned structure. Using the definitions given for S_n, Y_n, and τ_n given in the previous paragraphs, we will consider the Markov renewal process embedded at instants S_n just after service completion and defined by the Markov renewal sequence $\{(Y_n, \tau_n) : n \geq 0\}$. For $i, i' \geq 0$ and for $x \geq 0$, standard renewal arguments are used to obtain

$$Q^*_{ii'}(x) = \begin{cases} \int_0^x \lambda e^{-\lambda u} Q^*_{1,i'}(x-u)\, du, & i = 0,\ i' \geq 0, \\ \int_0^x e^{-\lambda u} \dfrac{(\lambda u)^{i'-i+1}}{(i'-i+1)!}\, dH(u), & i \geq 1,\ i' \geq i-1, \\ 0, & i' < i-1. \end{cases}$$

Define the following joint probabilities for $x > 0$ and $\nu = 0, 1, 2, \ldots$:

$$\begin{aligned} B_\nu(x) &= P\left\{\nu \text{ arrivals occur during } (0,\tau],\ \tau \leq x,\ |\, Y_0 = 0\right\}, \\ A_\nu(x) &= P\left\{\nu \text{ arrivals occur during } (0,\tau],\ \tau \leq x,\ |\, Y_0 > 0\right\}. \end{aligned}$$

It is evident from these definitions that

$$\begin{aligned} B_{i'}(x) &= Q^*_{0i'}(x), \quad i' \geq 0 \\ A_{i'-i+1}(x) &= Q^*_{ii'}(x), \quad i \geq 1,\ i' \geq 0. \end{aligned}$$

If we let x tend towards infinity and define

$$B_\nu = \lim_{x\to\infty} B_\nu(x), \qquad A_\nu = \lim_{x\to\infty} A_\nu(x),$$

we obtain the transition probability matrix $P = Q^*(\infty)$ of the embedded Markov chain, which assumes the following structure:

$$P = \begin{bmatrix} B_0 & B_1 & B_2 & B_3 & B_4 & \cdots \\ A_0 & A_1 & A_2 & A_3 & A_4 & \cdots \\ 0 & A_0 & A_1 & A_2 & A_3 & \cdots \\ 0 & 0 & A_0 & A_1 & A_2 & \cdots \\ 0 & 0 & 0 & A_0 & A_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \tag{3.37}$$

As we shall soon see, processes with embedded transition probability matrices of the *canonical* form (3.37) comprise a distinct class of stochastic processes with a common set of steady-state solution methods.

3.2.2 Level-Independent Processes

Stochastic processes whose embedded Markov chains exhibit the characteristic form (3.37) are termed *level-independent processes of $M/G/1$ type*. This is the simplest manifestation of the $M/G/1$-type process in that the terms A_i, $i = 1, 2, 3, \ldots$ do not vary based upon row membership. As the form of this matrix shows, the embedded Markov chain may transition across any number of columns to the right while being restricted to a single transition to the left. This is the reason why $M/G/1$-type processes are often referred to as *skip-free-to-the-left.* Although for the general $M/G/1$-type Markov chain it is possible to specify infinite-dimensional block terms A_ν and B_ν for $\nu \geq 0$, we shall assume henceforth that the blocks are finite with dimension m. Also, as with the definition of the QBD transition probability matrix, we denote the first coordinate of each state as the *level* and those that comprise the remainder, *phases*.

For the sake of completeness, we present here a dual class of processes of $GI/M/1$-*type* for which P takes the canonical form

$$P = \begin{bmatrix} C_0 & D_0 & 0 & 0 & 0 & \cdots \\ C_1 & D_1 & D_0 & 0 & 0 & \cdots \\ C_2 & D_2 & D_1 & D_0 & 0 & \cdots \\ C_3 & D_3 & D_2 & D_1 & D_0 & \cdots \\ C_4 & D_4 & D_3 & D_2 & D_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{3.38}$$

As in the case of processes of $M/G/1$-type, the structure of this matrix makes clear the origin of the term *skip-free-to-the-right.* The analysis of such systems is considerably easier than that of type $M/G/1$ due to the fact that the *matrix-geometric* property holds for the steady-state probabilities of system size. It is interesting to note that QBDs are skip-free-to-the-right and -to-the-left simultaneously, and are consequently of both $GI/M/1$- and $M/G/1$-type. In particular, the classification of QBDs as $GI/M/1$-type processes is consistent with the matrix-geometric nature of their steady-state probabilities.

The (steady-state) theory of $M/G/1$-type processes has been well-established for quite some time (see [82]) and applies to a large number of related systems. The next class of processes that we discuss does not possess, at least to the same degree, a generally-defined set of methods that allow for its easy solution. In what follows, we will review the properties of level-dependent $M/G/1$-type processes that are relevant to the determination of their stability and to the computation of their steady-state distributions. This will later be used in the formulation of a new condition for the stability of such systems in Chapter 5.

3.2.3 Level-Dependent Processes

A natural generalization of the concept of $M/G/1$-type processes is that in which the embedded transition probability matrix $[P_{ii'}]$ is dependent upon i, which is the previous number in system. In this case, we obtain the following form of the transition matrix:

$$P = \begin{bmatrix} B_0 & B_1 & B_2 & B_3 & B_4 & \cdots \\ A_0^{(1)} & A_1^{(1)} & A_2^{(1)} & A_3^{(1)} & A_4^{(1)} & \cdots \\ 0 & A_0^{(2)} & A_1^{(2)} & A_2^{(2)} & A_3^{(2)} & \cdots \\ 0 & 0 & A_0^{(3)} & A_1^{(3)} & A_2^{(3)} & \cdots \\ 0 & 0 & 0 & A_0^{(4)} & A_1^{(4)} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{3.39}$$

in which the distributions of the increments of the process vary depending upon the current number in system. We denote processes with embedded transition probability matrices of the form (3.39) *level-dependent.* Otherwise, we shall refer to the process as being *level-independent.* As the reader may recall, an analogous distinction was made in the case of QBDs, and which required a bit of effort in order to characterize the stability conditions that prevailed in the more-general level-dependent type of process.

3.2.4 Ergodicity Conditions

As hinted previously, there exists, for an embedded *level-independent* Markov chain of $M/G/1$-type, a well-developed theory of ergodicity, as evidenced by the many publications on the subject; see [8, 12, 13, 22, 67, 82, 89, 92], among others. What follows here is a review of basic, but essential facts concerning the stability and steady-state distribution of *irreducible* processes of $M/G/1$-type as it is presented in [82: Chap 2]. We first consider the criteria that determine the ergodicity of the embedded Markov chain $Q^*(\infty)$. As noted by Neuts [82] and others, it is important to recognize the fact that the ergodicity of the embedded chain is, in general, only a *necessary* condition for the recurrence of the associated Markov renewal process $Q^*(\cdot)$, as one must further stipulate that the mean increment of the MRP for any given time interval $x > 0$ is finite. However, as in the case of the QBD, the ergodicity of the embedded chain is also *sufficient* if the process has a finite number of phases; e.g., a process modulated by a random environment with a finite state space. Since we assume such a process in this dissertation, it becomes necessary only to ascertain the conditions that guarantee the ergodicity of the embedded chain.

Central to this concept of the ergodicity of an $M/G/1$-type Markov chain is the *fundamental period*, which was defined in Section 3.1.3 for the QBD. This quantity is of paramount importance to the determination of the recurrence properties of the states of any queueing system. Using our previous notation, let us now consider the probabilities $G(\nu, x) = G_1(\nu, x)$. We then have the following relationship between these matrices of mass-functions and the matrices $A_\nu(x)$ of transition probabilities, a result that appears in [82: p 80]:

Theorem 3.12. *The matrices $G(\nu, x)$ satisfy the difference equations*

$$G(1,x) = A_0(x), \quad G(\nu,x) = \sum_{k=1}^{\infty} A_k(x) * G_k(\nu-1,x), \quad \nu \geq 2, \tag{3.40}$$

where $(\cdot, \cdot, *)$ is the matrix convolution product defined in Appendix A, (1.4).

The preceding theorem makes rigorous the notion of the fundamental period as potentially including one or more periods between transition epochs of the embedded Markov chain.

We next define the familiar matrix $\hat{G}(x)$ of probability mass functions as

$$\hat{G}(x) = \sum_{\nu=0}^{\infty} \hat{G}(\nu, x), \quad x \geq 0, \tag{3.41}$$

which thus satisfies the relation $\hat{G}(x)\boldsymbol{e} \leq \boldsymbol{e}$. The matrix $\hat{G}(x)$ again denotes the *distribution matrix of the fundamental period.* As such, we may interpret $\hat{G} = \hat{G}(\infty)$ as the matrix whose (j, j')-th entry is the probability that the process will attain the state (i, j'), given that it started in $(i+1, j)$. Also, as shown in [82: Sec 2.3], $\hat{G}$ is the minimal nonnegative solution to the equation

$$\hat{G} = \sum_{\nu=0}^{\infty} A_\nu \hat{G}^\nu. \tag{3.42}$$

The following Lemma is clear from the probabilistic interpretation of $\hat{G}$ and the definition of recurrence:

Lemma 3.1. *If the $M/G/1$-type process with the embedded Markov chain whose transition probability matrix is given by $Q^* = Q^*(\infty)$ is recurrent, then the matrix $\hat{G} = \hat{G}(\infty)$ is stochastic.*

Thus, it is *necessary* that $\hat{G}$ be stochastic in order for the process be recurrent. Although we ultimately seek to define the necessary and sufficient criteria for positive recurrence to hold, it is first necessary to define an attainable analytic condition that is equivalent to the stochasticity of the matrix $\hat{G}$.

It is generally known from Perron-Frobenius theory that a given square matrix M is stochastic if and only if its maximum positive eigenvalue (i.e., its *Perron eigenvalue*) $sp(M)$ is equal to unity. A simple analytical condition that guarantees the fulfillment of $sp(M) = 1$ has been derived in [82: Chap 3]. In order to state this condition, we first define the matrix z-transform

$$A^*(z) = \sum_{\nu=0}^{\infty} A_\nu z^\nu$$

and

$$\beta = \frac{d}{dz} A^*(z)\bigg|_{z=1^-} = \sum_{\nu=0}^{\infty} \nu A_\nu,$$

where $|z| \leq 1$. Suppose as well that $\boldsymbol{\pi}$ is the invariant probability vector associated to the stochastic matrix $A = A^*(1^-)$. We may now state the following theorem (see [82: Thm 2.3.1]):

Theorem 3.13. *If the matrix A is irreducible, then the matrix G is stochastic if and only if*

$$\rho = \boldsymbol{\pi}\beta\boldsymbol{e} \leq 1. \tag{3.43}$$

Hence, the embedded Markov process $Q^* = Q^*(\infty)$ is recurrent if and only if inequality (3.43) *holds.*

One must keep in mind, however, that Theorem 3.13 only guarantees *recurrence* of the embedded Markov chain. Positive recurrence of a discrete-time Markov chain $\{X_n : n \geq 0\}$ requires that for any initial state i, $\xi_j = E[T_j \mid X_0 = i] < \infty$, where T_j is the first time to reach the state j, and $i \neq j$. Since the transition probability matrix of the embedded Markov chain $Q^*(\infty)$ is irreducible, the recurrence properties of every state is exactly the same, and so it suffices to verify this condition for the initial state 0. One may thus conclude that *positive recurrence* holds if and only if

$$\sum_{\nu=0}^{\infty} \nu B_\nu < \infty.$$

It is also shown in Section 3.2 of [82] that $\rho = 1$ corresponds to the null-recurrent case. Taken together, these facts lead us to the following conclusion:

Theorem 3.14. *Suppose that the matrix $\hat{G}$ that corresponds to the embedded Markov process $Q^*(\infty)$ is irreducible. Then the process $Q^*(\infty)$ is positive recurrent if and only if $\rho < 1$ and $\sum_{\nu=0}^{\infty} \nu B_\nu < \infty$.*

Since, by definition, the matrix β contains the conditional expectations of the number of arrivals during a fundamental period of the embedded Markov chain $Q^*(\infty)$, ρ may be interpreted as the expected number of arrivals during a fundamental period. It is intuitive that if this number exceeds unity, then the queue will tend to grow in the long run. This interpretation is essential to the formulation of the stability criteria for level-dependent $M/G/1$-type queues.

This chapter has reviewed the fundamental concepts and definitions required for the steady-state analysis of quasi-birth-and-death and $M/G/1$-type systems. The material presented here is crucial to understanding the main results obtained in this dissertation which are presented in the next two chapters. The next topic is the formulation and steady-state analysis of the unreliable $M/M/1$ retrial queue in a random environment, for which the material in Section 3.1 shall be relevant. Once the results for the Markovian model have been established, then it will be shown that a similar queueing model with *generally* distributed service requirements is a process of $M/G/1$-type. This will subsequently allow the use of the theory and methods described in Section 3.2. Results for system stability have likewise been extended from the level-independent results of this chapter and will appear in the subsequent chapters.

4. Exponential Service Requirements

This chapter provides the mathematical model description and main results for a single-server unreliable retrial queue in a random environment with exponential services and Poisson arrival process. We first characterize the system as a *quasi-birth-and-death* (QBD) process, which facilitates a matrix-analytic approach. Once this has been accomplished, the theory of quasi-birth-and-death processes may be brought to bear on the tasks of (1) determining the conditions under which a steady-state distribution of orbit size exists, and (2) the computation of the approximate steady-state distribution and related performance measures.

4.1 Model Description and Notation

Consider a single-server M/M/1 retrial queue in which the model parameters (arrival, service, failure, repair, and retrial rates) are all modulated by an external random environment (see Figure 4.1). The random environment is assumed to be an irreducible continuous-time Markov chain (CTMC) with a finite state space $S = \{1, \dots, m\}$, infinitesimal generator $Q = [q_{ij}]_{i,j \in \{1,\dots,m\}}$ and stationary probability vector $\boldsymbol{p} = [p_1, \dots, p_m]$. When the environment is in state j, customers arrive to the system according to a Poisson process with rate $\lambda_j > 0$ while their service requirements are exponentially distributed with rate $\mu_j > 0$. When the server is either idle or busy, breakdowns occur according to a Poisson process with rate $\xi_j > 0$, and the subsequent exponential repair time has rate $\alpha_j > 0$. Define the m-dimensional vectors $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_m)$, $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_m)$, $\boldsymbol{\xi} = (\xi_1, \xi_2, \dots, \xi_m)$, and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_m)$. Retrial customers attempt to regain access to the server independently of all other customers at exponentially distributed time intervals with rate $\theta_j > 0$ when the environment is in state j. The vector of retrial rates is $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_m)$. For a row vector $\boldsymbol{x}$, we define the diagonal matrix of its elements by $\Delta(\boldsymbol{x}) = diag(\boldsymbol{x})$. Arrival, service, failure, repair, and retrial processes are assumed

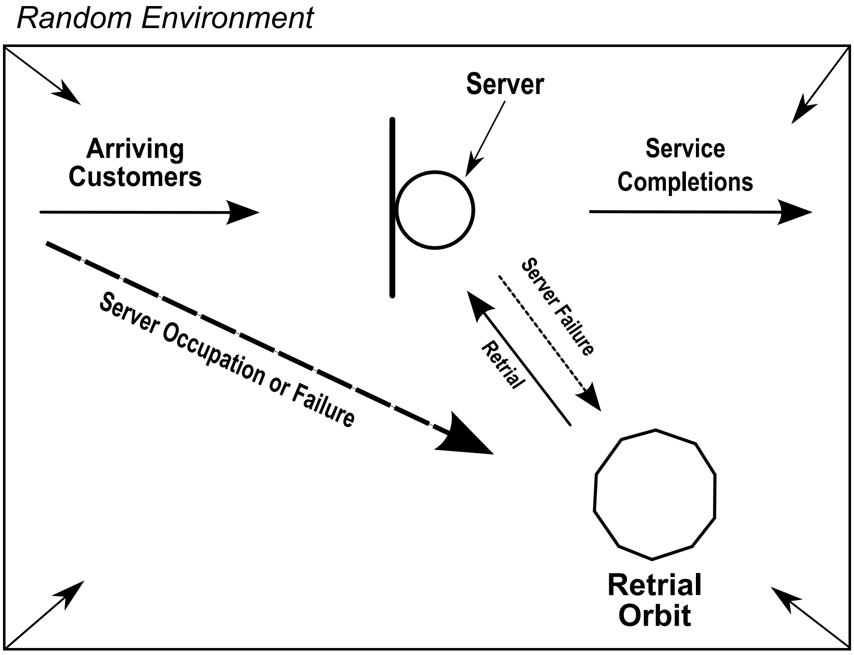

Figure 4.1 Graphical depiction of a single-server retrial queue.

to be mutually independent. Denote by $R(t)$, the number of customers in orbit at time t and let $Z(t)$ be the state of the random environment at time t. The random variable $X(t)$ denotes the status of the server defined by

$$X(t) = \begin{cases} 0 & \text{if the server is failed at time } t \\ 1 & \text{if the server is operational but idle at time } t \\ 2 & \text{if the server is operational and busy at time } t \end{cases} .$$

The continuous-time stochastic process, $\{(R(t), Z(t), X(t)) : t \geq 0\}$, has state space $\mathcal{S} = \{\,(i, j, k) : i \in \mathbb{Z}^+, j \in S, k \in \{0, 1, 2\}\,\}$, where $\mathbb{Z}^+$ is the set of nonnegative integers. It is clear from the foregoing definitions that the process is a multivariate Markov chain with one infinite dimension, namely the number of customers in orbit. Next, we show that $\{(R(t), Z(t), X(t)) : t \geq 0\}$ can be viewed as a tri-layered QBD and exploit this structure to obtain the stability condition for the queueing system and to obtain the approximate steady-state distribution and mean performance measures.

Let us denote the infinitesimal generator of $\{(R(t), Z(t), X(t)) : t \geq 0\}$ by Q^*. Figure 4.2 shows scalar elements of the infinitesimal generator are arranged in $m \times m$ diagonal sub-blocks corresponding to the m states of the random environment. In this figure, we are restricted to displaying the entries of Q^* when $m = 3$. Using the Δ-terminology to denote diagonal matrices, we aggregate terms and rewrite the *level-1* generator in block form as follows:

$$Q_1^* = \begin{bmatrix} C_0 & \Delta(\boldsymbol{\xi}) & \Delta(\boldsymbol{\lambda}) & 0 & 0 & 0 & 0 & \cdots \\ \Delta(\boldsymbol{\alpha}) & D_1 & 0 & 0 & \Delta(\boldsymbol{\lambda}) & 0 & 0 & \cdots \\ \Delta(\boldsymbol{\mu}) & 0 & D_2 & 0 & \Delta(\boldsymbol{\xi}) & \Delta(\boldsymbol{\lambda}) & 0 & \cdots \\ 0 & 0 & \Delta(\boldsymbol{\theta}) & C_1 & \Delta(\boldsymbol{\xi}) & \Delta(\boldsymbol{\lambda}) & 0 & \cdots \\ 0 & 0 & 0 & \Delta(\boldsymbol{\alpha}) & D_1 & 0 & 0 & \cdots \\ 0 & 0 & 0 & \Delta(\boldsymbol{\mu}) & 0 & D_2 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{4.1}$$

where the matrices D_1, D_2, and C_i, $i \geq 1$ are given by

$$\begin{aligned} D_1 &\equiv Q - \Delta(\boldsymbol{\lambda} + \boldsymbol{\alpha}), \\ D_2 &\equiv Q - \Delta(\boldsymbol{\lambda} + \boldsymbol{\mu} + \boldsymbol{\xi}), \\ C_i &\equiv \Delta(\boldsymbol{\mu}) + D_2 - i\Delta(\boldsymbol{\theta}), \quad i = 1, 2, \ldots. \end{aligned}$$

(R,Z,X)	$(0,1,1)$	$(0,2,1)$	$(0,3,1)$	$(0,1,0)$	$(0,2,0)$	$(0,3,0)$	$(0,1,2)$	$(0,2,2)$	$(0,3,2)$	$(1,1,1)$	$(1,2,1)$	$(1,3,1)$	$(1,1,0)$	$(1,2,0)$	$(1,3,0)$	$(1,1,2)$	$(1,2,2)$	$(1,3,2)$	$\cdots$
$(0,1,1)$	${d_{011}}^\dagger$	q_{12}	q_{13}	ξ_1	0	0	λ_1	0	0	0	0	0	0	0	0	0	0	0	$\cdots$
$(0,2,1)$	q_{21}	d_{021}	q_{23}	0	ξ_2	0	0	λ_2	0	0	0	0	0	0	0	0	0	0	$\cdots$
$(0,3,1)$	q_{31}	q_{32}	d_{031}	0	0	ξ_3	0	0	λ_3	0	0	0	0	0	0	0	0	0	$\cdots$
$(0,1,0)$	α_1	0	0	d_{010}	q_{12}	q_{13}	0	0	0	0	0	0	λ_1	0	0	0	0	0	$\cdots$
$(0,2,0)$	0	α_2	0	q_{21}	d_{020}	q_{23}	0	0	0	0	0	0	0	λ_2	0	0	0	0	$\cdots$
$(0,3,0)$	0	0	α_3	q_{31}	q_{32}	d_{030}	0	0	0	0	0	0	0	0	λ_3	0	0	0	$\cdots$
$(0,1,2)$	μ_1	0	0	0	0	0	d_{012}	q_{12}	q_{13}	0	0	0	ξ_1	0	0	λ_1	0	0	$\cdots$
$(0,2,2)$	0	μ_2	0	0	0	0	q_{21}	d_{022}	q_{23}	0	0	0	0	ξ_2	0	0	λ_2	0	$\cdots$
$(0,3,2)$	0	0	μ_3	0	0	0	q_{31}	q_{32}	d_{022}	0	0	0	0	0	ξ_3	0	0	λ_3	$\cdots$
$(1,1,1)$	0	0	0	0	0	0	θ_1	0	0	d_{111}	q_{12}	q_{13}	ξ_1	0	0	λ_1	0	0	$\cdots$
$(1,2,1)$	0	0	0	0	0	0	0	θ_2	0	q_{21}	d_{121}	q_{23}	0	ξ_2	0	0	λ_2	0	$\cdots$
$(1,3,1)$	0	0	0	0	0	0	0	0	θ_3	q_{31}	q_{32}	d_{131}	0	0	ξ_3	0	0	λ_3	$\cdots$
$(1,1,0)$	0	0	0	0	0	0	0	0	0	α_1	0	0	d_{110}	q_{12}	q_{13}	0	0	0	$\cdots$
$(1,2,0)$	0	0	0	0	0	0	0	0	0	0	α_2	0	q_{21}	d_{120}	q_{23}	0	0	0	$\cdots$
$(1,3,0)$	0	0	0	0	0	0	0	0	0	0	0	α_3	q_{31}	q_{32}	d_{130}	0	0	0	$\cdots$
$(1,1,2)$	0	0	0	0	0	0	0	0	0	μ_1	0	0	0	0	0	d_{112}	q_{12}	q_{13}	$\cdots$
$(1,2,2)$	0	0	0	0	0	0	0	0	0	0	μ_2	0	0	0	0	q_{21}	d_{122}	q_{23}	$\cdots$
$(1,3,2)$	0	0	0	0	0	0	0	0	0	0	0	μ_3	0	0	0	q_{31}	q_{32}	d_{132}	$\cdots$
$(2,1,1)$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	$2\theta_1$	0	0	$\cdots$
$(2,2,1)$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	$2\theta_2$	0	$\cdots$
$(2,1,1)$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	$2\theta_3$	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

$^\dagger d_{ijk}$ is the negative of the sum of all other elements along the row corresponding to state (i,j,k).

Figure 4.2 The infinitesimal generator matrix Q^*.

Visual inspection of (4.1) does not help us discern if Q^* is a QBD. However, we observe that the diagonal blocks repeat in cycles of length three, namely $\{C_i,\, D_1,\, D_2\}$ as $i \to \infty$. This suggests a higher-order aggregation of 3×3 block entries (or $3m \times 3m$ scalar entries), from which we obtain the top-level (0) matrix form for the infinitesimal generator Q_0^*:

$$Q_0^* = \begin{bmatrix} \Gamma_0 & \Lambda & 0 & 0 & 0 & \cdots \\ \Theta_1 & \Gamma_1 & \Lambda & 0 & 0 & \cdots \\ 0 & \Theta_2 & \Gamma_2 & \Lambda & 0 & \cdots \\ 0 & 0 & \Theta_3 & \Gamma_3 & \Lambda & \cdots \\ 0 & 0 & 0 & \Theta_4 & \Gamma_4 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{4.2}$$

where, for $i \geq 0$,

$$\Gamma_i \equiv \begin{bmatrix} C_i & \Delta(\boldsymbol{\xi}) & \Delta(\boldsymbol{\lambda}) \\ \Delta(\boldsymbol{\alpha}) & D_1 & 0 \\ \Delta(\boldsymbol{\mu}) & 0 & D_2 \end{bmatrix}, \tag{4.3}$$

$$\Theta_i \equiv \begin{bmatrix} 0 & 0 & i\Delta(\boldsymbol{\theta}) \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tag{4.4}$$

$$\Lambda \equiv \begin{bmatrix} 0 & 0 & 0 \\ 0 & \Delta(\boldsymbol{\lambda}) & 0 \\ 0 & \Delta(\boldsymbol{\xi}) & \Delta(\boldsymbol{\lambda}) \end{bmatrix}. \tag{4.5}$$

At this level of grouping, it is clear that Q^* possesses the tridiagonal form of a QBD, and thus, the following theorem may be stated:

Theorem 4.1. *The process* $\{(R(t), Z(t), X(t)) : t \geq 0\}$ *is a tri-layered, level-dependent QBD.*

As of the writing of this dissertation, known results for multi-layered and level-dependent QBDs do not provide much guidance in determining stability conditions for the retrial queueing model $\{(R(t), Z(t), X(t))\}$. Consequently, such theoretical conditions must be established before an attempt to quantify the performance measures of the system can be made. In the next section, we give conditions for the stability of *general* level-dependent, discrete-time QBDs. What is more, we will show that these conditions depend upon the QBD structure, and not upon the number of layers that it encompasses.

4.2 Stability and Steady-State Analysis

We now set out to prescribe the analytic conditions to guarantee the ergodicity of the unreliable retrial queueing system in a random environment. In what follows, we present the formulation and subsequent proof of the conditions for the ergodicity of general discrete-time level-dependent QBDs. The finite cardinality of the phase spaces (namely the environment and server states) corresponding to each level i allows the use of the discrete embedded chain at transitions of the original continuous-time QBD process in determining conditions on the model parameters that guarantee its stability. An explicit traffic intensity formula ρ, which does not depend upon the initial level i, will then be derived for the $M/M/1$ model described in this chapter.

As its name suggests, a (discrete-time) level-dependent QBD possesses drift values that vary with the level of the process. This introduces a major difficulty in that the nonstationary nature of the increments of the random walk $\mathcal{L}_n = L_{k_n}$ render the epochs $k_1, k_2, k_3, \ldots$ of the return of the process $\{\delta_n\}$ to δ_0 non-regenerative. Nevertheless, we shall show that there indeed exists a criterion based upon the level-dependent drift $d(i)$ that enables the determination of the stability of the QBD. Before presenting this criterion, however, we will re-introduce the random-walk process and associated terminology as they pertain to a general level-dependent QBD.

We begin as before with an analysis of the random walk $L_n = L_0 + \sum_{k=1}^{n} \delta_k$ at steady-state, where each term is defined as in (3.26). The definition of the processes W_L and W_δ and the corresponding limiting increment δ are identical. This time, however, the drift $d(i)$ at each level $i \geq 1$, where $d(i) \equiv E[\, \delta_n \,|\, L_{n-1} = i]$, varies according to the current level i. We may thus consider $d(i)$ to be the long-run drift of the process restricted to level i, $i = 0, 1, 2, \ldots$. Contrast this definition to the analogous one for level-independent QBDs, whose homogeneity of drift over levels allows us to state the correspondence $d(i) = d(j) = \bar{d}$ for all i, $j \geq 0$.

For a discrete-time, level-independent QBD with a finite number of phases, we have seen that the related Markov chain W_L has the finite steady-state distribution

$$(\boldsymbol{\alpha} A_2 \boldsymbol{e}, \boldsymbol{\alpha} A_1 \boldsymbol{e}, \boldsymbol{\alpha} A_0 \boldsymbol{e}),$$

but its level-dependent counterpart

$$(\boldsymbol{\alpha}^{(i)} A_2^{(i)} \boldsymbol{e}, \boldsymbol{\alpha}^{(i)} A_1^{(i)} \boldsymbol{e}, \boldsymbol{\alpha}^{(i)} A_0^{(i)} \boldsymbol{e}), \quad i = 0, 1, 2, \ldots,$$

is now infinite-dimensional. Nevertheless, just as in the level-independent case, we may evaluate the drift associated with level i, $d(i)$ as

$$d(i) \;=\; \boldsymbol{\alpha}^{(i)} A_0^{(i)} \boldsymbol{e} - \boldsymbol{\alpha}^{(i)} A_2^{(i)} \boldsymbol{e}, \quad i = 0, 1, 2, \ldots, \tag{4.6}$$

with $\boldsymbol{\alpha}^{(i)}$ being the minimal positive solution to the system of equations given by

$$\boldsymbol{\alpha}^{(i)} A^{(i)} \;=\; \boldsymbol{\alpha}^{(i)}, \quad \boldsymbol{\alpha}^{(i)} \boldsymbol{e} \;=\; 1, \qquad i = 0, 1, 2, \ldots$$

and $A^{(i)} \equiv A_0^{(i)} + A_1^{(i)} + A_2^{(i)}$ is the transition probability matrix of the QBD restricted to level i. Note that, by virtue of the dependence of the drift upon level index i, that Theorem 3.9 does not apply to the level-dependent case. Despite this, we may still establish a single condition for ergodicity based upon the convergence of the sequence $\{d(i)\}$ as i grows without bound. The sufficiency of this condition is provided by Pakes' Lemma. However, the much more difficult task of demonstrating necessity shall require the use of Theorem 3.7, after which we shall state the corresponding result for continuous-time, level-dependent QBDs.

It is now apparent that we require a suitable embedded Markov chain. By observing that the sequence $\{(\mathcal{L}_n, S_n) : n \geq 0\}$, which is embedded at level-jump epochs S_n such that $\mathcal{L}_n = X^+_{S_n}$, is a Markov renewal sequence, we show that $\{\mathcal{L}_n\}$ is a Markov chain.

Lemma 4.1. *Let $\{(X_n, Y_n) : n \geq 0\}$ be a discrete-time, level-dependent QBD, and let S_n denote the time of the nth level jump with $S_0 = 0$. Define $\mathcal{L}_n = X^+_{S_n}$, the level of the QBD just after S_n. Then $\{\mathcal{L}_n : n \geq 0\}$ is a Markov chain.*

Proof. We shall first demonstrate that the increments of $\mathcal{L}$ are PH-distributed (see Section A.2 of the Appendix). From previous discussions, it is clear that the matrix $A_1^{(i)}$, $i \geq 0$, contains the probabilities for transitions *within* a given level i, while $A_0^{(i)} + A_2^{(i)}$ contains those for transitions *out* of level i conditioned upon any of the states $S_\phi^{(i)} = \{1, 2, \ldots, M\}$ in the phase process $\{\phi_n^{(i)} : n \geq 0\}$ for level i. We may thus represent the time the process $\mathcal{L}$ spends between transitions as the time until absorption, where the absorbing state shall be denoted by 0 and the conditional probabilities of absorption are given by the vector $(A_0^{(i)} + A_2^{(i)})\boldsymbol{e}$. Thus, the inter-transition times are PH-distributed with the representation $[\boldsymbol{\gamma}_\ell^{(i)}, A_1^{(i)}]$, where $\boldsymbol{\gamma}_\ell^{(i)}$ is an initial probability vector.

As a consequence of the PH-distributions of increments $\mathcal{L}_{i+1} - \mathcal{L}_i$ according to parameters that reside only within a particular level i, the Markov property holds over such increments, and thus

$$\{(\mathcal{L}_n, S_n) : n \geq 0\}$$

is a Markov renewal sequence. Thus, by Theorem 3.11, we have that $\{\mathcal{L}_n\}$ is a Markov chain. □

Finally, we define certain quantities that are needed for the proof of our main theorem. The first is simply a restatement of the drift terms $d(i)$ as ratios, or expressions of *traffic intensity*; see [83] for a detailed discussion of the traffic intensity $\rho^{(i)}$.

$$\rho^{(i)} \;=\; \frac{\boldsymbol{\alpha}^{(i)} A_0^{(i)} \boldsymbol{e}}{\boldsymbol{\alpha}^{(i)} A_2^{(i)} \boldsymbol{e}}, \qquad i = 0, 1, 2, \ldots. \tag{4.7}$$

In order to make this expression well-defined, we will exclude the case in which one or more members of the sequence $\{A_2^{(i)}\}$ has a zero eigenvector. It is easy to see that (4.7) and the drift at level i are related by the expression

$$d(i) \;=\; (\rho^{(i)} - 1)(\boldsymbol{\alpha}^{(i)} A_2^{(i)} \boldsymbol{e}) \qquad i = 0, 1, 2, \ldots. \tag{4.8}$$

In order to simplify certain expressions that contain these terms, we define $\gamma_j^{(i)} = \boldsymbol{\alpha}^{(i)} A_j^{(i)} \boldsymbol{e}, \quad i, j = 0, 1, 2,$ which will allow us to then rewrite (4.8) in terms of the traffic intensity $\rho^{(i)}$ as

$$d(i) \;=\; (\rho^{(i)} - 1)\gamma_2^{(i)}, \qquad i = 0, 1, 2, \ldots. \tag{4.9}$$

We may now state and prove the following theorem.

Theorem 4.2. *An irreducible, aperiodic, discrete-time, level dependent QBD is ergodic if and only if*

$$\rho \;\equiv\; \lim_{i \to \infty} \rho^{(i)} \;<\; 1, \tag{4.10}$$

with $\rho^{(i)}$ defined as in (4.7).

Proof. We first assume that inequality (4.10) holds. By (4.9) we obtain

$$\limsup_{i\to\infty} d(i) \;=\; (\rho - 1)\limsup_{i\to\infty} \gamma_2^{(i)}. \tag{4.11}$$

Assume that $\rho < 1$. Since it is impossible to have an increment (a jump between levels) of magnitude greater than unity in a QBD, the finite-increment stipulation of Pakes' Lemma (c.f. [87] and [61: pp 96–97]) is fulfilled. Next, we observe that since $\gamma_2^{(i)} > 0$ for all i, we likewise obtain that $\limsup_{i\to\infty} \gamma_2^{(i)} > 0$. Consequently,

$$\limsup_{i\to\infty} d(i) \;=\; (\rho - 1)\limsup_{i\to\infty} \gamma_2^{(i)} \;<\; 0.$$

With the remaining condition of Pakes' Lemma satisfied, we may conclude that the QBD is positive recurrent if $\rho < 1$.

The proof of necessity hinges upon the fulfillment by the QBD of Kaplan's condition (see Section 3.1.3). Let $z \in [0, 1)$ and define the functions

$$\psi_i(z) \;=\; \left(z^i - \sum_{j\geq 1} p_{ij} z^j \right) / (1 - z), \quad i \geq 0 \tag{4.12}$$

as in (3.25). It was shown in Section 3.1.3 that the equivalence $d(i) \;=\; \lim_{z\to 1^-} \psi_i(z)$ holds, and thus demonstrates the relationship of the function $\psi_i(z)$ to the drift condition of the Markov chain. It is also mentioned that the fulfillment of Kaplan's condition implies that the drift quantities $d(i), i \geq 0$ are bounded below. If, in addition, we have that $d(i) \geq 0$ for every i past a certain level N, then it becomes clear from this discussion that the Markov chain cannot be ergodic; this is the reasoning behind Theorem 1 of [95], which we shall show holds under the hypothesis of our theorem. In doing so, we thus prove the contrapositive of the statement of necessity, which will complete our proof.

Assume that $\rho \geq 1$ and that the system is stable; we need to show that the QBD cannot be ergodic. Using the same reasoning as in the proof of sufficiency, we assert that $d(i) < \infty$ for each $i \geq 0$. In order to prove that Kaplan's condition holds, we will derive the function

$$\psi_i(z) \;=\; \left(z^i - \sum_{j \geq 1} p_{ij} z^j \right) / (1-z), \quad i \geq 0, \; z \in [0,1)$$

specifically for the Markov chain $\{(\delta_n, \phi_n) : n \geq 0\}$ consisting of the increments of the random walk L_n (3.26). Thus, δ_n may assume any one of the three values $\{-1, 0, +1\}$ with the corresponding steady-state probabilities $[\,\boldsymbol{\gamma}^{(0)}, \boldsymbol{\gamma}^{(1)}, \boldsymbol{\gamma}^{(2)}, \ldots]$, where $\boldsymbol{\gamma}^{(i)} = [\,\gamma_2^{(i)}, \gamma_1^{(i)}, \gamma_0^{(i)}]$. This results in the following expression:

$$\psi_i(z) \;=\; -\frac{z^{i-1}}{1-z}\,[\,\gamma_2^{(i)} + (\gamma_1^{(i)} - 1)z + \gamma_0^{(i)} z^2], \quad i \geq 1. \tag{4.13}$$

We next investigate the behavior of the sequence of terms $\{\psi_i(\hat{z}_i) : i = 2, 3, \ldots\}$ evaluated at the critical points $\hat{z}_i$ for which ψ_i achieves a local minimum over the interval $(0,1)$. After some tedious algebra, we obtain

$$\hat{z}_i \;=\; \left(1 - \frac{1}{i}\right) \frac{\gamma_2^{(i)}}{\gamma_0^{(i)}} \;=\; \left(1 - \frac{1}{i}\right) \frac{1}{\rho^{(i)}}, \tag{4.14}$$

and thus,

$$\psi_i(\hat{z}_i) \;=\; \left(1 - \frac{1}{i}\right)^i \gamma_2^{(i)} \left[1 - \frac{i}{i-1}\right]. \tag{4.15}$$

Because the probability of downward drift increases with the level, it is seen that $\{\gamma_2^{(i)}\}$ is nondecreasing, and thus, so is $\{\psi_i(\hat{z}_i)\}$. Consequently, it may be readily verified that $\psi_i(\hat{z}_i) \geq 0$ for $i \geq 2$. We observe that, for each $i \geq 2$, $\psi_i(z)$ is nondecreasing in z on the interval $[\,\hat{z}_i, 1)$. This, together with the continuity of ψ_i over the interval $[\,0,1)$ further implies that, for each $i \geq 2$, $\psi_i(z) \geq 0$ for all $z \in [\,\hat{z}_i, 1)$. Finally, we observe that $\{1/\rho^{(i)}\}$ is a nondecreasing sequence, which is also due to the fact that drift decreases as the level increases due to the geometric rate of decay of the steady-state probabilities in an ergodic Markov chain; thus, $\{\hat{z}_i\}$ is nondecreasing and approaches $1/\rho$ as a limit. Thus, we conclude that $\psi_i(z) \geq 0$ for all

$z \in [1/\rho, 1)$ and for each $i \geq 2$. Hence, Kaplan's condition is satisfied. To prove the existence of a positive integer N such that $d(i) \geq 0$ for every $i \geq N$, we observe that, for each $\varepsilon > 0$, we have $\Delta\rho = |\rho^{(i)} - \rho| \leq \varepsilon$ whenever $i \geq N_\varepsilon$ for some $N_\varepsilon > 0$. By our hypothesis, $\rho > 1$, so we set $\varepsilon = \rho - 1$. Removing the absolute value in $\Delta\rho$, we obtain

$$1 - \rho \ \leq \ \rho^{(i)} - \rho \ \leq \ \rho - 1, \quad i \geq N_{\rho-1}, \tag{4.16}$$

which shows that $\rho^{(i)} \geq 1$ for $i \geq N_{\rho-1}$. By equation (4.8), we now see that $d(i) \geq 0$ for $i \geq N_{\rho-1}$, and thus, by Theorem 1 of [95], we conclude that the QBD is not positive recurrent, and hence not ergodic. Therefore, it is necessary that $\rho < 1$ for ergodicity. □

Although Theorem 4.2 applies to discrete-time systems, it may be easily extended to the continuous-time version using an embedded discrete-time Markov chain. By considering the Markov chain embedded at level jump epochs in our unreliable retrial queueing model, we directly apply Theorem 4.2 to obtain the condition required to ensure stability of the queueing system.

Theorem 4.3. *An irreducible, aperiodic, continuous-time, level-dependent QBD is ergodic if and only if*

$$\rho \ \equiv \ \lim_{i\to\infty} \rho^{(i)} \ < \ 1, \tag{4.17}$$

with $\boldsymbol{\alpha}^{(i)}$ defined as the unique positive solution to the system of equations given by

$$\boldsymbol{\alpha}^{(i)} A^{(i)} \ = \ 0, \quad \boldsymbol{\alpha}^{(i)} \boldsymbol{e} \ = \ 1, \quad i = 0, 1, 2, \ldots,$$

where $A^{(i)} \equiv A_0^{(i)} + A_1^{(i)} + A_2^{(i)}$ and $\boldsymbol{e}$ is a column vector containing ones.

Proof. This assertion is a direct consequence of Theorem 4.2 and Theorem 7.2.4 of [66] (page 158). □

Theorem 4.3 gives us the means by which we may characterize the stability for the queueing process $\{(R(t), Z(t), X(t) : t \geq 0)\}$.

Theorem 4.4. *The QBD with infinitesimal generator matrix Q^* is positive recurrent if and only if*

$$\rho \equiv \frac{-\boldsymbol{e}'\,[\,(Q-\Delta(\boldsymbol{\alpha})\,)\Delta(\boldsymbol{\xi}) \;+\; (Q-\Delta(\boldsymbol{\alpha}+\boldsymbol{\xi})\,)\Delta(\boldsymbol{\lambda})\,]\;D_\alpha^{-1}\,\boldsymbol{e}}{\boldsymbol{e}'\,(Q-\Delta(\boldsymbol{\alpha})\,)\,(Q-\Delta(\boldsymbol{\mu}+\boldsymbol{\xi})\,)\;D_\alpha^{-1}\,\boldsymbol{e}} \;<\; 1, \tag{4.18}$$

where $D_\alpha \equiv Q - \Delta(\boldsymbol{\alpha}+\boldsymbol{\xi})$ and $\boldsymbol{e}$ is a column vector containing ones as entries.

Proof. We begin with the derivation of the sequence $\{\rho^{(i)} : i = 1, 2, \ldots\}$ for the QBD with infinitesimal generator Q^*. In order to facilitate the computation of the limit over i, we purposely exclude the degenerate situation in which $\boldsymbol{\alpha}^{(i)}A_2^{(i)}\boldsymbol{e} = 0$, thus making the sequence $\{\rho^{(i)}\}$ well-defined for each i. It is now possible to apply (4.7) to the elements of the QBD defined in (4.2), upon which we obtain

$$\rho^{(i)} \;=\; \frac{-\boldsymbol{e}'\Delta(\boldsymbol{\lambda})\,\Delta(\boldsymbol{\xi}) + (Q-\Delta(\boldsymbol{\alpha})\,)(i\Delta(\boldsymbol{\theta})+\Delta(\boldsymbol{\lambda})\,)\Delta(\boldsymbol{\lambda}+\boldsymbol{\xi})\,D_{i\theta}^{-1}\,)D_\alpha^{-1}\,\boldsymbol{e}}{\boldsymbol{e}'\,i\Delta(\boldsymbol{\theta})\,(Q-\Delta(\boldsymbol{\alpha})\,)(Q-\Delta(\boldsymbol{\mu}+\boldsymbol{\xi})\,)D_{i\theta}^{-1}D_\alpha^{-1}\,\boldsymbol{e}}, \tag{4.19}$$

where $D_{i\theta}^{-1} \;=\; (Q - i\Delta(\boldsymbol{\theta}) - \Delta(\boldsymbol{\lambda}+\boldsymbol{\mu}+\boldsymbol{\xi})\,)$. Using standard techniques for the determination of limits of rational functions, the limit of (4.19) as $i \to \infty$ gives us (4.18), as desired. Note that this result implies the existence of the limit for all choice of parameters, excluding those that result in the degenerate case discussed at the beginning of this proof. □

When there is only a single environment state, our queueing model corresponds to the model of Kulkarni and Choi [62] with exponentially distributed service, interfailure, and repair times. In such case, the expression for ρ in (4.18) reduces to Equation (2.4) of [62].

4.3 Performance Measures

Approximate performance measures for the queueing system $\{(R(t), Z(t), X(t)) : t \geq 0\}$ in steady-state may be obtained with numerical results for the steady-state distribution, which is given by the vector $\boldsymbol{\pi} = [\pi(i,j,k)]$, where

$$\pi(i,j,k) \;=\; \lim_{t\to\infty} P\left\{R(t) = i, Z(t) = j, X(t) = k\right\}, \quad (i,j,k) \in \mathcal{S}.$$

We begin with a discussion of the expected size of the retrial orbit R in steady-state, which is given by

$$E[R] \;=\; \sum_{i=1}^{\infty} i\,(\boldsymbol{\pi}_i \boldsymbol{e}). \tag{4.20}$$

Since the steady-state probabilities correspond to each state of the trivariate representation of the queueing process, we must aggregate these probabilities in such a way as to obtain the probability of being in orbit; in other words, for the sequence $\{p_i : i \geq 0\}$ of steady-state probabilities of the orbit size i, we obtain $p_i \;=\; \boldsymbol{\pi}_i \boldsymbol{e}$.

Many other performance measures may be obtained in much the same way as was done for the expected size of the orbit. For instance, let us consider the limiting proportion of time that the system is under repair. For each orbit size i, we must find all possible phases for which $X \;=\; 0$; in this case, the long-run proportion of time that the system is down is given by

$$P_0 \equiv P(\text{System is Down}) \;=\; \sum_{i=0}^{\infty}\sum_{j=1}^{m} \pi(i,j,0). \tag{4.21}$$

We may therefore conclude that the proportion of time the server spends in state $k,\ k = 0,1,2$, is given by

$$P_k \;=\; \sum_{i=0}^{\infty}\sum_{j=1}^{m} \pi(i,j,k), \qquad k = 0,1,2. \tag{4.22}$$

The expected number in system $E[L]$ and system sojourn time $E[W]$ follow directly from the orbit-size probabilities and according to the following logic. We base our approach to the computation of $E[L]$ upon the observation that the system size always equals the orbit size whenever the system is down or the server is idle.

When the server is busy, one is added to the expected number $E[R]$ in orbit. Thus, we have

$$\begin{aligned} E[L] &= E[R]\,(P_0 + P_1) + (E[R]+1)P_2 \\ &= E[R] + P_2. \end{aligned}$$

We next turn our attention to the computation of $E[W]$. For work-conserving systems, the typical approach is to employ Little's Law, which relates $E[L]$ to $E[W]$. It is given by the well-known formula

$$E[L] \;=\; \bar{\lambda}E[W], \quad E[R] \;=\; \bar{\lambda}E[W_R],$$

where $\bar{\lambda}$ is the *average* rate of arrivals to the system in steady state. Given the equilibrium probability vector $\boldsymbol{p} = [p_j : j = 1, \ldots, m]$ of the random environment, we obtain the average arrival rate $\bar{\lambda} = \boldsymbol{p\lambda' e}$, and then solve for $E[W]$:

$$E[W] \;=\; E[L]/\bar{\lambda}, \qquad E[W_R] \;=\; E[R]/\bar{\lambda}. \tag{4.23}$$

4.4 Useful Algorithms

As a result of the recursive elements of many of its defining analytical results, QBD methods do not readily admit closed-form solutions for the steady-state distribution. There do exist certain classes of *level-independent* QBD processes for which an explicit rate matrix R may be derived; for this, we refer the reader to [108]. Nevertheless, the model that we study in this chapter is not amenable to the methods described in [108], and so we must resort to algorithmic techniques for the computation of an approximate stationary distribution. In this section, we shall review the algorithms of Bright and Taylor [25] and discuss their implementation. In the section that follows, we shall compute some important measures of interest using the approximated steady-state distribution.

In discussing the algorithms of this section, we shall refer to the continuous-time level-dependent QBD (LDQBD) as the bivariate process $\{(X(t), Y(t)) : t \geq 0\}$. Nevertheless, it should be clear to the reader that we may substitute for this a multi-layered (i.e. n-variate for $n \geq 3$) LDQBD process if so desired. It should also be mentioned that the more general case of the LDQBD, in which the number of phases M_i corresponding to level i for $i = 0, 1, 2, \ldots$ may vary, is assumed in [25]. We will state Theorem 4.2 with this assumption intact. However, we will henceforth assume that $M = M_i = M_j < \infty$ for every $i, j \in \mathbb{Z}^+$; in other words, the QBD will be homogeneous in the number of phases at each level.

It has already been mentioned that the existence of a limiting distribution hinges upon the behavior of the QBD at level 0. It is therefore fortunate that the process restricted to level 0 of a continuous-time QBD is itself a CTMC that possesses a time-to-absorbtion (transition to the next level) that is PH-distributed with representation $(\boldsymbol{\alpha}^{(0)}, Q^{(0)})$, where we define $Q^{(0)} \equiv A_1^{(0)} + A_0^{(0)} G^{(1)}$, and $\boldsymbol{\alpha}^{(0)}$ is the positive vector solution to the system

$$\boldsymbol{\alpha}^{(0)} Q^{(0)} = 0, \qquad \sum_{j=0}^{M} \alpha_j^{(0)} = 1.$$

As always, we assume that there are $M < \infty$ states (phases) in level 0 and, furthermore, it is clear that $Q^{(0)}$ is the infinitesimal generator of the restricted process.

The statement of Theorem 3.4 asserts the necessary and sufficient conditions for the ergodicity of a level-dependent QBD process. However, setting these criteria to the task of obtaining a numerical steady-state distribution entails the solution of a system of linear equations for *each level* i, which in turn would require explicitly-defined $G^{(i)}$ matrices. Since QBDs that possess such explicitly-defined G matrices are few and far between, this would, at best, be deemed an untenable approach. Nevertheless, as in [25], Theorem 3.4 may serve as the basis for efficient algorithms dedicated to obtaining the steady-state distributions of level-dependent QBDs.

The conditions for the existence of a stationary distribution of a continuous-time level-dependent QBD, which is that the vector of stationary probabilities $\boldsymbol{\pi}$ be the unique solution to the system of equations given by (3.17) and (3.18), clearly shows that the rate matrix $R^{(i)}$, $i = 0, 1, 2, \ldots$, is key to its determination. The algorithms developed by Bright and Taylor [25] are well-suited to the purpose of computing $\boldsymbol{\pi}$ using standard mathematical computing environments such as MATLAB®, as we have done for the unreliable retrial queue in a random environment. The fundamental idea behind this computational procedure is based upon the following result, which is given as Lemma 1 in [25]:

Lemma 4.2. *(Bright and Taylor) If the level-dependent QBD given by $\{(X(t), Y(t)) : t \geq 0\}$ with state space $\{ (i, j) : i \geq 0,\ 1 \leq j \leq M_j\}$ is positive recurrent, then the sequence $\{R_i : i = 0, 1, 2, \ldots\}$ is given by*

$$R_i = \sum_{k=0}^{\infty} U_i^k \prod_{r=0}^{k-1} D_{i+2^{r-k}}^{r-1-k}, \quad i \geq 0, \tag{4.24}$$

where U_i^r and D_i^r are $M_{i-1} \times M_{i-1+2^r}$ and $M_{i-1} \times M_{i-1-2^r}$ matrices respectively and are given by the recursive expressions

$$\begin{aligned}
U_i^0 &= A_0^{(i)}(-A_1^{(i+1)})^{-1} \quad \textit{for } k \geq 1, && (4.25)\\
D_i^0 &= A_2^{(i)}(-A_1^{(i-1)})^{-1} \quad \textit{for } k \geq 1, && (4.26)\\
U_i^{r+1} &= U_i^r U_{i+2^r}^r [I - U_{i+2^{r+1}}^r D_{i+3\cdot 2^r}^r - D_{i+2^{r+1}}^r U_{i+2^r}^r]^{-1}, && (4.27)\\
D_i^{r+1} &= D_i^r D_{i-2^r}^r [I - U_{i-2^{r+1}}^r D_{i-2^r}^r - D_{i-2^{r+1}}^r U_{i-3\cdot 2^r}^r]^{-1}. && (4.28)
\end{aligned}$$

Lemma 4.2 is the focal point of a series of four algorithms whose overall objective is to compute an approximate steady-state distribution. One may, of course, repeatedly apply Lemma 4.2 to find each rate matrix, but this is not necessary due to the following relationship:

$$A_0^{(i)} + R_i A_1^{(i+1)} + R_i[R_{i+1} A_2^{(i+2)}] \; = \; 0, \quad i \geq 0, \tag{4.29}$$

which allows R_i to be defined recursively as

$$R_i \; = \; A_0^{(i)}[-A_1^{(i+1)} - R_{i+1} A_2^{(i+2)}]^{-1}. \tag{4.30}$$

Thus, it is more efficient to obtain an approximation for R_{K-1} for some threshold level K and then work backwards using equation (4.30) to obtain all of the remaining rate matrices. Then, after solving the system in Theorem 3.1 to determine $\boldsymbol{\pi}_0(K)$, the remaining vectors $\boldsymbol{\pi}_i(K), \; i > 0$ may be obtained through successive applications of equation (3.5) while normalizing the set of probabilities with each iteration i. This notion provides the basis for Algorithm 1 in Figure 4.3, which is due to Bright and Taylor [25]. This defines the overall algorithm that defines our method of approaching the computation of the approximate steady-state probabilities. Several of the key steps of this algorithm are addressed in further detail as sub-algorithms, each of which likewise appeared in the article by Bright and Taylor [25].

Step (2) of Algorithm 1 is rather involved, as it requires an iterative application of Lemma 4.2. The purpose of the algorithm is to obtain an approximation for R_{K-1} based upon some predefined tolerance level $\epsilon > 0$ that dictates the number of terms of the sequence $\{R_{K-1}(N) : N \geq 1\}$ defined by

$$R_i(N) \equiv \sum_{k=0}^{N} U_i^k \prod_{r=0}^{k-1} D_{i+2^{r-k}}^{r-1-k}, \quad i, N \geq 0 \tag{4.31}$$

that we must obtain for a given threshold $K > 0$. In other words, $R_i(N)$ is the $N+1$-th partial sum of (4.24). We thus iterate through successive values of $R_{K-1}(N), \; N \geq 1$ until

$$|\, R_{K-1}(N) - R_{K-1}(N-1) \,|_\infty \; < \; \epsilon, \tag{4.32}$$

Algorithm 1 (Bright and Taylor [25]):

```
1. Determine threshold level $K$.
2. Obtain $R_{K-1}$ using Algorithm 2.
3. Obtain the remaining $R_i,\ 0 \le i \le K-2$ using (4.29).
4. Solve the following for $\boldsymbol{\pi}_0(K)$:
       $\boldsymbol{\pi}_0(K)(A_1^{(0)} + R_0 A_2^{(1)}), \quad \boldsymbol{\pi}_0(K)\boldsymbol{e} = 1.$
5. for $i = 1 : K$
       Compute $\boldsymbol{\pi}_i(K) = \boldsymbol{\pi}_{i-1}(K) R_{i-1}$.
       Normalize $\{\boldsymbol{\pi}_k(K) : k = 0, 1, \ldots, K\}$ so that $\sum_{k=0}^{K} \boldsymbol{\pi}_k = 1$.
```

Figure 4.3 Top-level algorithm for computing the steady-state distribution $\{\boldsymbol{\pi}_i(K) : 0 \le i \le K-1\}$.

where $|A|_\infty = \max_{i,j}(A_{ij})$, $A \in \mathbb{R}^{m \times n}$ is the $\boldsymbol{L}^\infty$ norm. This results in Algorithm 2, which is depicted in Figure 4.4. Since Algorithm 2 involves the computation of the U_i^k and $D_{i+2^{k+1}}^k$ matrices, which are requisite for obtaining the N-th partial sum $R_{K-1}(N)$, we must define yet another subprocedure, namely Algorithm 3, in order to wholly specify the procedure given in Algorithm 2. We shall discuss the elaboration of this next task in the paragraph that follows.

For convenience, we now define the *UD-pair* $UD(N, i)$ as the ordered pair $(U_i^k, D_{i+2^{k+1}}^k)$ whose constituent matrices appear in Step (2) of Algorithm 2. Based on the recursive definitions of these matrices in (4.25)–(4.28), it might seem that an algorithm to compute the UD-pair need only consist of their repeated evaluation, and, as such, is straightforward in execution. However, implementing these recursive relations directly will most likely result in the repeated computation of the same UD-pairs, which is clearly undesirable. The algorithm obtains and stores only those pairs $UD(\eta, \iota)$ that are necessary to complete the recursion up to $UD(N, i)$, $N \geq$

Algorithm 2 (Bright and Taylor [25]):

```
1. Initialize the following variables:
       Set $N = 0$.                                   } Steps 1 and 2 of Algorithm 3.
       Set $U = U_i^0,\ D = D_{i+2}^0$.               }
       Set $\Pi = I$ (Identity matrix).
       Set $R_i(0) = U$.
2. Increment $N$: $N = N + 1$.
3. Make the following assignments:
       $\Pi = D \cdot \Pi$.
       $U = U_i^N, \quad D = D_{i+2^{N+1}}^N$.        } Obtained from Algorithm 3
       $R_i(N) = R_i(N-1) + U \cdot \Pi$.
4. If $|R_i(N) - R_i(N-1)| \geq \epsilon$, then go back to step (2).
5. Set $R = R_i(N)$.
```

Figure 4.4 Procedure to compute the rate matrix R_{K-1} in step (2) of Algorithm 1.

0. Fortunately, the relationships among UD-pairs are regular, and thus may be represented pictorially as in [25: Fig 2]. The algorithm itself appears as Figure 4.5.

Finally, and perhaps most critically, is the question of how to choose the maximum level K so as to ensure the validity of the approximate steady-state probability vector $\boldsymbol{\pi}(K)$. The obvious criterion would be to set K large enough so that the sum of all steady-state probabilities is close to 1. Putting this idea into practice, however, is difficult if one desires an *a priori* answer – which is essential if one wishes to avoid the brute-force alternative of running the algorithm until the desired accuracy is achieved. What further makes this idea so unattractive is the explosion of the number of UD-pairs – and hence matrix inverses – that one needs to compute as the values of K increase. This situation is readily apparent if one observes the rapid

Algorithm 3 (Bright and Taylor [25]):

```
1. if $N = 0$ then compute $UD(0, i)$ and store.
2. for $m = 0$ to $N$
       for $n = K + (2^{N-n+1} - 1)2^i$
           to  $K + (2^{N-n+1} - 1)2^{i+1}$ step $2^n$
               Evaluate $UD(m, n)$ and store.
3. Increment: $N = N + 1$.
4. Using all $UD$-pairs stored thus far, compute $UD(N, i)$ and store it as
   well.
5. Remove all stored $UD$-pairs except $UD(n, i + (2^{N-m} - 1)2^{j+1})$, $j =
   0, 1, \ldots, N$.
6. Return to Step 2 if Algorithm 2 has not yet terminated.
```

Figure 4.5 Procedure to compute $UD(N, i)$ for the Algorithm 2 (c.f. Fig. 4.4).

growth of the number of UD-pairs contained within the outer trapezoidal region of the graph in [25: Fig 2]. In order to counter this difficulty, a heuristic of sorts is presented in the form of Algorithm 4 in [25]. Simply put, it is a reformulation of Algorithm 1 with the option to increase the value of K until $\boldsymbol{\pi}_K \boldsymbol{e} < \epsilon$, for some arbitrarily chosen $\epsilon > 0$. The algorithm is given in Figure 4.6.

As noted in [25], it is theoretically possible to have a level-dependent QBD whose $A_k^{(i)}$ matrices are not parametrically defined (that is, explicitly dependent upon level $i \geq 0$ and a finite number of parameters). In this case, the application of the algorithms discussed in this section requires the input of a large number of unrelated $A_k^{(i)}$ matrices into the computer in order to reach the required tolerance for the smallest steady-state probability. Even if this were done, however, the unpredictability of the QBD generator blocks for a nonparametric QBD would make this

Algorithm 4 (Bright and Taylor [25]):

```
1. Set $K_{prev} = 0$ and choose some tolerance $\epsilon > 0$.
2. Set $K_{new} > K_{prev}$.
3. Compute the partial sum $R_{K_{new}-1}(4)$ of expression (4.24).
4. if $|R_i(4) - R_i(3)| \geq \epsilon$ then go back to Step (2).
5. Compute $R_{K_{new}-2}, R_{K_{new}-3}, \ldots, R_{K_{prev}}$ using the recursive definition
   (4.30) of the rate matrix $R_i$.
6. Solve the system
       $\boldsymbol{\pi}_0(K_{new})(A_1^{(0)} + R_0 A_2^{(1)}) = 0, \quad \boldsymbol{\pi}_0(K_{new})\boldsymbol{e} = 1.$
7. for $i = K_{prev} + 1$ to $K_{new}$
       $\boldsymbol{\pi}_i(K_{new}) = \boldsymbol{\pi}_{i-1}(K_{new}) R_{i-1}$.
       Normalize $\boldsymbol{\pi}_0(K_{new}), \boldsymbol{\pi}_1(K_{new}), \ldots, \boldsymbol{\pi}_i(K_{new})$
           so that $\sum_{k=0}^{i} \boldsymbol{\pi}_k(K_{new})\boldsymbol{e} = 1$.
8. Set $K_{prev} = K_{new}$.
9. if $\boldsymbol{\pi}_{K_{new}}(K_{new})\boldsymbol{e} \geq \epsilon$ then go back to Step (2).
10. Set $K = K_{prev}$.
```

Figure 4.6 Procedure to compute $\boldsymbol{\pi}(K)$ for some suitably-chosen $K > 0$.

determination tricky at best. It is therefore recommended in [25] that, even in the parametric case, one should take great care in determining stopping criteria when applying the algorithms.

4.5 Busy Period Analysis

We approach the study of the busy period of the system, as in Section 3.1.3, in terms of the *fundamental period* of the system. For the QBD process of this chapter, we may formally express the distribution function $G^{\circ}(\tau)$ of the busy period as the $m \times m$ matrix

$$[G^{\circ}(\tau)]_{kk'} = P\{\tau < \infty,\ (Y(\tau), Z(\tau), X(\tau)) = (0, 1, k') \\ \ldots \,|\, (Y(0), Z(0), X(0)) = (1, 2, k)\},$$

where $k, k' \in \{1, \ldots, m\}$,

$$Y(t) = \begin{cases} R(t) + 1 & \text{if } X(t) = 2 \\ R(t) & \text{otherwise.} \end{cases} \quad \text{and} \quad (R(t), Z(t), X(t)) \in \mathcal{S}.$$

A complicating issue is that failures of the server may preclude the servicing of customers in a busy period; in addition, the server of a retrial queue may be up and idle with customers still in orbit, which is likewise before the termination of the busy period. Hence, it is more instructive to study the fundamental period corresponding to level i, whose distribution (matrix) function is denoted by $[G^{(i)}(x)]_{jj' \in S},\ i \in \mathbb{Z}^{+}$.

Application of the block matrix components (4.3) through (4.5) of the QBD representation of the queue $\{(R(t), Z(t), X(t)) : t \geq 0\}$ to Theorem 3.10 results in a matrix quadratic equation which, when solved, gives the transform $\widetilde{G}^{(i)}(z, s)$ of the matrix distribution of the fundamental period. For the purpose of simplification, let

$$\chi_i(z, s) \ = \ \widetilde{G}^{(i)}(z, s).$$

We thus obtain the equation

$$\chi_i(z, s) = \Big(i\Delta(\boldsymbol{\theta})\, C_1 + \chi_i^2(z, s) C_2^{(i)}\Big)\, \Upsilon_i^{-1}, \tag{4.33}$$

where, if we define $T_\lambda = Q - \Delta(\boldsymbol{\lambda})$,

$$C_1 = \begin{bmatrix} 0 & 0 & (\Delta(\boldsymbol{\alpha}) - T_\lambda)\,(\Delta(\boldsymbol{\xi} + \boldsymbol{\mu}) - T_\lambda) \\ 0 & 0 & \Delta(\boldsymbol{\alpha})\,(\Delta(\boldsymbol{\xi} + \boldsymbol{\mu}) - T_\lambda) \\ 0 & 0 & \Delta(\boldsymbol{\mu})\,(\Delta(\boldsymbol{\alpha}) - T_\lambda) \end{bmatrix}$$

$$C_2^{(i)} = \left[\begin{array}{lll} 0 & \Delta(\boldsymbol{\xi})\,\Delta(\boldsymbol{\lambda})\,(\Delta(\boldsymbol{\xi} + \boldsymbol{\mu} + \boldsymbol{\alpha}) - 2T_\lambda) & \ldots \\ 0 & \Delta(\boldsymbol{\lambda})\,\{T_\lambda^2 - [\Delta(\boldsymbol{i\theta}) + \Delta(\boldsymbol{\xi})\,\Delta(\boldsymbol{\mu} + \boldsymbol{e})]\,T_\lambda & \\ & \qquad + [\Delta(\boldsymbol{\xi} + \boldsymbol{i\theta}) + \Delta(\boldsymbol{\xi})\,\Delta(\boldsymbol{\alpha}) - \Delta(\boldsymbol{\lambda})\,\Delta(\boldsymbol{\mu})]\} & \ldots \\ 0 & \Delta(\boldsymbol{\xi})\,\{T_\lambda^2 - \Delta(\boldsymbol{\alpha} + \boldsymbol{\xi} + \boldsymbol{i\theta}) + \Delta(\boldsymbol{\lambda})\,\Delta(\boldsymbol{\mu})\} & \ldots \end{array}\right.$$

$$\left.\begin{array}{lr} \ldots & \Delta(\boldsymbol{\lambda})^2\,(\Delta(\boldsymbol{\alpha}) - T_\lambda) \\ \ldots & \Delta(\boldsymbol{\lambda})^2\,\Delta(\boldsymbol{\alpha}) \\ \ldots & \Delta(\boldsymbol{\lambda})\,[T_\lambda^2 - \Delta(\boldsymbol{\alpha} + \boldsymbol{\xi} + \boldsymbol{i\theta})\,T_\lambda + \Delta(\boldsymbol{i\theta})\,\Delta(\boldsymbol{\alpha})] \end{array}\right]$$

for $i \geq 0$, and

$$\begin{aligned} \Upsilon_i \;=\; & -T_\lambda^3 + \Delta(\boldsymbol{\alpha}{+}\boldsymbol{2\xi}{+}\boldsymbol{\mu}{+}\boldsymbol{i\theta})\,T_\lambda^2 \\ & \qquad - [\Delta(\boldsymbol{\alpha}{+}\boldsymbol{\xi}{+}\boldsymbol{i\theta})\,\Delta(\boldsymbol{\xi}{+}\boldsymbol{\mu}) - \Delta(\boldsymbol{\lambda})\,\Delta(\boldsymbol{\mu})]\,T_\lambda \\ & \qquad\qquad + \Delta(\boldsymbol{\alpha})\,[i\Delta(\boldsymbol{\theta})\,\Delta(\boldsymbol{\xi}{+}\boldsymbol{\mu}) - \Delta(\boldsymbol{\lambda})\,\Delta(\boldsymbol{\mu})]\,. \end{aligned}$$

The matrix G of the fundamental period is thus the smallest positive solution to (4.33). The necessary and sufficient conditions for the (exact) solvability of matrix-quadratic equations may be found in [98]. However, due to the complexity of the foregoing expression, a closed-form expression for the transform may not be derived, and so we must resort to an approximation using either the simple algorithm (3.36) or by obtaining the matrices through computation of the steady-state distribution via the algorithms that were presented in Section 4.4.

4.6 Numerical Illustrations

We conclude this chapter with numerical illustrations of the results we have obtained for the queueing system $\{(R(t), Z(t), X(t)) : t \geq 0\}$. Note that, in addition to the overall traffic intensity $\rho = \lim_{i \to \infty} \rho^{(i)}$ which is derived from Theorem 4.2, we may compute the traffic intensity ρ_j for the system when the environment is in state j as follows:

$$\rho_j = \frac{\lambda_j \xi_j + \alpha_j(\lambda_j + \xi_j)}{\alpha_j(\mu_j + \xi_j)}, \qquad j \in S. \tag{4.34}$$

The traffic intensity (4.34) may be obtained through simplification of the traffic intensity model of [62] or by fixing the parameters of the model for $m \geq 2$ and then applying (4.7) and (4.10).

4.6.1 Example 1: Three-State Environment

We first consider a system operating in a three-state random environment whose infinitesimal generator is given by

$$Q = \begin{bmatrix} -1.0 & 1.0 & 0.0 \\ 1.6 & -2.6 & 1.0 \\ 0.0 & 3.2 & -3.2 \end{bmatrix}. \tag{4.35}$$

Table 4.1 summarizes the system parameter values. Notice that the traffic intensity for state 1 of the environment exceeds unity, and, thus, the system may experience periods of instability whenever the environment is in state 1. Nevertheless, the overall traffic intensity obtained by (4.7) and (4.10), is $\rho = 0.4730 < 1$, so it is (overall) a stable system.

Table 4.2 lists the first 20 equilibrium probabilities of orbit size. A graphical depiction of these probabilities – together with an additional 56 probabilities – is shown in Figure 4.7. The probabilities represented in this figure were computed

Table 4.1 Summary of parameters for 3-state example.

Environment State (j)	λ_j	μ_j	ξ_j	α_j	θ_j	ρ_j
1	1.0	0.5	9.0	4.0	1.5	1.2895
2	0.1	0.8	2.0	7.0	3.0	0.7602
3	3.0	10.0	0.1	2.0	5.0	0.3218

Table 4.2 Steady-state probabilities of orbit size for Example 1.

i	p_i	i	p_i
0	0.02798110	11	0.03540854
1	0.03306333	12	0.03376359
2	0.03706997	13	0.03208472
3	0.03973863	14	0.03039734
4	0.04131836	15	0.02872157
5	0.04205032	16	0.02707325
6	0.04212703	17	0.02546462
7	0.04169928	18	0.02390498
8	0.04088593	19	0.02240120
9	0.03846185	20	0.02095816
10	0.03698781		

using $\sum_{j=1}^{3}\sum_{k=0}^{2}\pi(i,j,k)$ for each $i \in \{0,1,2,\ldots\}$. Note the geometric rate of decay exhibited by the steady-state distribution, which is a defining characteristic of an ergodic system.

The efficient recursive algorithms of Bright and Taylor [25], are useful for computing the steady state probabilities of level-independent QBDs and may be adapted to our model. Here we make use of Algorithms 1 through 4 of [25] to obtain (approximate) steady-state probabilities from which we may compute queueing performance measures. Let $\pi_{i,j,k}(t) = P(R(t)=i, Z(t)=j, X(t)=k)$ and define as before

$$\pi(i,j,k) = \lim_{t\to\infty}\pi_{i,j,k}(t), \qquad (i,j,k)\in\mathcal{S}.$$

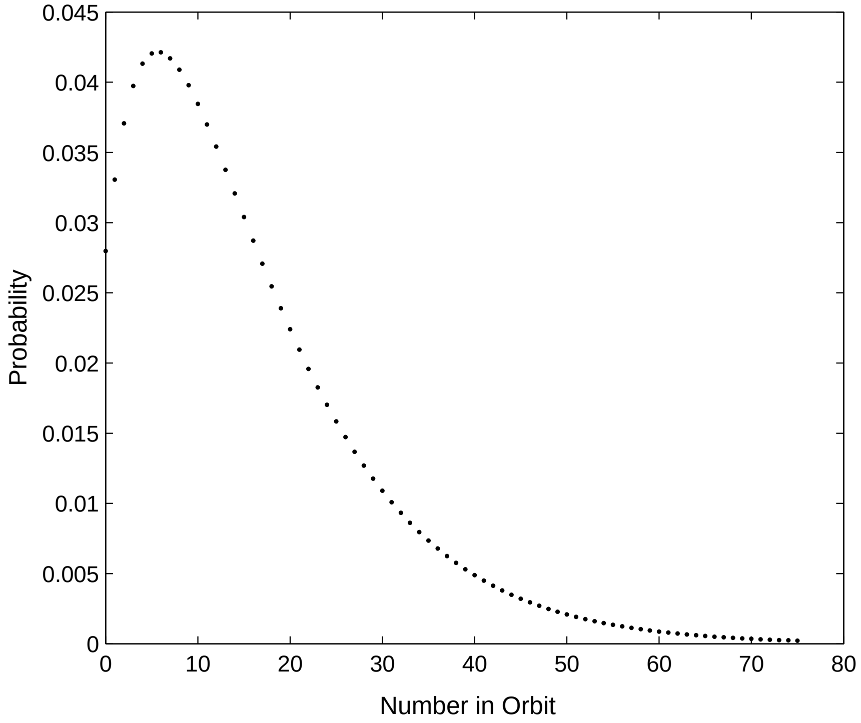

Figure 4.7 Pictorial representation of steady-state orbit-size probabilities for Example 1.

The limiting probability that i customers are in the orbit is given by

$$\pi_i \equiv \sum_{j=1}^{3} \sum_{k=0}^{2} \pi(i,j,k), \quad i \geq 0.$$

The steady-state expected number of customers in the orbit ($E[R]$) is approximated (using the first 76 probabilities) by $E[R] = \sum_{i=0}^{75} i\, \pi_i \approx 15.6684$. We obtain the overall system size ($E[L]$) by summing the expected number in orbit and the expected number in service, namely, $E[L] = E[R] + P_2 \approx 16.0921$, where $P_2 = \sum_{i=0}^{75} \sum_{j=1}^{3} \pi(i,j,2) \approx 0.4237$ is the long-run probability that the server is not failed and busy. Applying Little's law and the average arrival rate ($\bar{\lambda} = \boldsymbol{\lambda p} \approx 0.90558$), we respectively obtain the expected time spent in the retrial orbit and in the system by

$$E[W_R] = E[R]/\bar{\lambda} \approx 17.302, \quad E[W] \; = \; E[L]/\bar{\lambda} \approx 17.77.$$

4.6.2 Example 2: Seven-State Environment

We shall next consider a 7-state example that exhibits interesting non-intuitive behavior as compared to the previous 3-state example. As before, we define the environment by its infinitesimal generator One might observe that the rates of transition *to* states 2 and 3 are larger (in general) than that of any of the other transitions. From the table of system parameter values shown in Table 4.3, we may further observe that the queue is subject to heavy traffic when the environment is in states 2 and 3. The values of these parameters were set in order to observe the effects of heavy traffic upon the stability of the system.

$$Q = \begin{bmatrix} -7.6 & 2.0 & 3.0 & 1.0 & 1.0 & 0.1 & 0.5 \\ 0.5 & -8.0 & 2.5 & 1.0 & 2.0 & 0.8 & 1.2 \\ 0.3 & 1.5 & -0.8 & 1.0 & 1.0 & 1.0 & 1.0 \\ 2.0 & 3.0 & 5.0 & -1.2 & 0.1 & 1.0 & 0.1 \\ 0.8 & 2.5 & 2.0 & 1.1 & -7.9 & 0.7 & 0.8 \\ 1.5 & 1.0 & 1.6 & 1.2 & 0.5 & -6.3 & 0.5 \\ 2.0 & 2.5 & 2.0 & 1.0 & 1.8 & 0.9 & -10.2 \end{bmatrix}. \tag{4.36}$$

Next, through the employment of the algorithms of Bright and Taylor, we computed the steady-state probabilities corresponding to the retrial queueing system thus defined, from which we subsequently obtain the probabilities of orbit size (c.f. Table 4.4).

The values of the various system performance measures is shown in Table 4.5. From this we can see that, although the greater stability of the system with a 7-state environment is not reflected in orbit or system sizes, it is definitely present in the

Table 4.3 Parameters for the system of Example 2.

Environment State j	Parameter					ρ_j
	λ_j	μ_j	ξ_j	α_j	θ_j	
1	3.0	7.0	0.5	2.0	1.0	0.5667
2	1.0	3.0	1.1	0.5	11.0	1.0488
3	3.0	12.5	1.5	0.5	2.0	0.9643
4	2.0	12.5	4.0	2.0	2.0	0.6061
5	2.0	4.5	1.0	6.0	5.0	0.6061
6	0.5	2.0	0.7	3.0	1.0	0.4877
7	0.5	4.0	1.5	0.5	0.5	0.6364
					ρ	0.1915

Table 4.4 Steady-state probabilities for orbit sizes up to and including $i = 20$ for the model presented in Example 2.

i	p_i	i	p_i
0	0.01470817	11	0.03366217
1	0.01887110	12	0.03322658
2	0.02262905	13	0.03261428
3	0.02577557	14	0.03185676
4	0.02832568	15	0.03098201
5	0.03032339	16	0.03001475
6	0.03182163	17	0.02897667
7	0.03287514	18	0.02788673
8	0.03353730	19	0.02676134
9	0.03385857	20	0.02561467
10	0.03388586		

Table 4.5 Performance measures for Examples 1 and 2.

Model	Performance Measure				R_{max}	ρ
	$E[L_R]$	$E[L]$	$E[W_R]$	$E[W]$		
3-State	15.668	16.092	17.302	17.770	7	0.4730
7-State	19.550	19.825	7.680	7.788	10	0.1915

comparison between sojourn times. The values for times in orbit and in system are less than half of what they are for the system with a three-state environment. The interpretation that one may ascribe to this behavior is that system size is kept in check – despite heavy traffic – by the relatively short times that the average customer spends in orbit and in service.

4.6.3 Comparison to Simulated Data

It was mentioned at the beginning of this chapter that the effective times of service, arrivals, breakdowns, repairs, and retrials will significantly affect how one models the retrial queueing system. Indeed, one must effectively assign PH-distributed durations to these activities, the capability of which is not inherently present in the graphical process simulator Arena®. As a consequence, two basic work-arounds need to be employed depending upon whether the activity is a "service"-type activity or an "arrival"-type activity. Let us first discuss the procedure for simulating the modulation of arrivals to the system. Suppose that the environment has undergone a transition from state i to state j. Simply changing the arrival rate will not work since all scheduled arrivals at the previous rate need to be instantiated before the new rate takes effect. This results in a backlog of customers from the arrival distribution of the previous environmental state, thus invalidating the statistics for the process under the new environmental state. It turns out that, in order to deal with instantaneous rate changes for arrivals, it is necessary to set the exponential arrival rate to a constant $\lambda_{max} = \max_{i \in \{1,\ldots m\}} (\lambda_i)$, and then allow only the proportion λ_j/λ_{max} of customers to enter the system.

Services, on the other hand, must be dealt with in a different manner since a customer in service has already been instantiated, as opposed to arriving customers whose *future* instantiation is affected by the exponential rate of arrival. These customers must have their current service preempted and they must then be reassigned a full exponential service time of rate μ_j (due to the memorylessness of the exponential distribution). Breakdowns are treated as arrivals to the system while repairs are dealt with in exactly the same way as services. The distribution of retrials is perhaps the most difficult to simulate. If there are k customers in orbit during the environmental transition, then there are two ways to proceed. One may either (1) pick an estimated maximum orbit size k_{max} and allow the ratio θ_j/θ_{max} of customers from an exponential retrial rate of $k_{max}\theta_{max}$ to proceed, or (2) preempt every customer in the retrial orbit and reassign the appropriate exponential inter-retrial duration. Each method has its difficulties, but (2) may be the preferred option for moderate-size environments due to the lack of the need to estimate a maximum orbit size, thereby saving memory and eliminating the need to produce large numbers of unneeded entities.

We will compare the average performance measures of 250 replications of Arena simulation output to the corresponding steady-state performance measures predicted by QBD approximation methods. The input of the simulated retrial queueing system is a Markov-modulated Poisson arrival process with services and breakdowns that are likewise modulated by the same three-state Markov chain whose infinitesimal generator is given by

$$Q = \begin{bmatrix} -0.2 & 0.05 & 0.15 \\ 0.3 & -0.4 & 0.1 \\ 0.01 & 0.7 & -0.71 \end{bmatrix}. \tag{4.37}$$

Using vector terminology, we may describe the exponential rates of the modulated parameters of the simulated model as $\boldsymbol{\lambda} = [0.3, 0.1, 3]$, $\boldsymbol{\mu} = [0.5, 0.8, 10]$, and $\boldsymbol{\xi} = [0.3, 1.5, 2.1]$. The exponential rates $\boldsymbol{\alpha}$ of repair and $\boldsymbol{\theta}$ of retrial were assigned the same rates over all environment states; that is, $\alpha_i = \alpha = 2.0$ and $\theta_i = \theta = 1.0$ for each $i \in S = \{1, 2, 3\}$. Applying the values of these parameters to (4.18) results in a value for the traffic-intensity factor ρ of 0.6630, which implies a stable, albeit relatively high-traffic system.

Table 4.6 displays the output of the simulation runs versus the corresponding measures computed from the steady-state system probabilities, which were, in turn, computed using the Bright and Taylor algorithm.

Table 4.6 Simulation output versus QBD approximation.

Performance Measure	Simulated	QBD Approximation
$E[\,L\,]$	7.5678	7.5784
$E[W]$	11.6288	11.6346
$E[\,R\,]$	7.0551	7.0645
$E[W_R]$	10.8402	10.8456
$P(\text{Busy})$	0.5127	0.5140
$P(\text{Failure})$	0.2920	0.2947

The results of this chapter enable us to characterize the explicit stability conditions and steady-state distribution of the exponential model using the matrix-analytic method. In addition, we have demonstrated the utility of matrix-geometric algorithms in approximating the steady-state probabilities, together with numerical results for the performance measures of the system. To this end, a simple criterion for the ergodicity of *general* level-dependent QBDs was proved, thus contributing to the understanding of the structure and behavior of this class of stochastic models. In the next chapter, a similar set of results will be established for a model with generally-distributed service requirements. The analysis of the steady-state behavior of the Markov-modulated $M/G/1$ retrial queue with unreliable server will have important implications, not only for the purpose of extending results to non-Markovian unreliable retrial queues, but also for elucidating the common elements that it shares with the Markovian model.

5. General Service Requirements

This chapter extends the results of Chapter 4 to investigate an unreliable retrial queue in a random environment wherein customers bring generally distributed service requirements. Following a brief model description, we explicitly derive the semi-Markov kernel corresponding to the Markov renewal sequence of orbit size and environmental state considered just after the system returns to an up-and-idle state. This will subsequently allow us to state specific conditions for the ergodicity of the embedded Markov chain. The matrix-analytic procedure that we employ makes significant use of the theory of Markov-modulated Poisson processes (MMPPs) whose rudimentary concepts are summarized in Section A.3 of the Appendix.

5.1 Model Description and Notation

We define the retrial queueing system in a very similar manner to that of the Markovian model, with the primary difference being the way in which the service-time distribution is characterized. We again define the continuous-time, trivariate-process description

$$\{(R(t), Z(t), X(t)) : t \geq 0\},$$

where $R(t) \in \mathbb{Z}^+$ denotes the orbit size at time t, $Z(t) \in S = \{1, 2, \ldots, m\}$ denotes the environmental state, and $X(t)$ the server status (0 = failed, 1 = idle, and 2 = busy). The arrival, retrial, failure, and repair processes are modulated by the Markov chain $\{Z(t) : t \geq 0\}$ with environment-dependent rates again contained in the vectors $\boldsymbol{\lambda}$, $\boldsymbol{\theta}$, $\boldsymbol{\xi}$, and $\boldsymbol{\alpha}$, respectively. The environment modulates the exponential rate parameters of the input streams via the evolution of its own state transitions. In other words, if the environment is in state $k \in S$, then the rate parameters are, respectively, λ_k, θ_k, ξ_k, and α_k.

Customers bring a service requirement represented by a nonnegative random variable with a continuous distribution function that is nondefective and proper.[1] Define the c.d.f. of service time as $H(t)$ with finite mean $1/\mu$ and denote its Laplace-Stieltjes transform (LST) by

$$\widetilde{H}(s) = \int_0^\infty e^{-st} dH(t).$$

We shall assume that the service is *not* modulated by the random environment, although this assumption can be relaxed for the case in which service times are distributed as PH-random variables (see Appendix, Section A.2.1).

Customers arrive according to a modulated Poisson process, are serviced by a single server, and then leave the system. Failures likewise occur as a Poisson process, upon which the server is occupied for the entire length of an exponentially-distributed repair period. A customer already in service when a failure occurs immediately proceeds to the retrial orbit, from whence it attempts to reaccess the server at random (exponential) intervals. Customers who arrive to a failed or busy server likewise join the retrial orbit. The retrials themselves are conducted independently of all others and succeed only when the server is up and idle. Departures from the system, however, only occur when a customer is not interrupted by a failure during the entire course of its service.

A major difference in the analysis of this system comes about as a result of the Markov-modulated input streams (otherwise known as Markov-modulated Poisson processes, or MMPPs; see Appendix, Section A.3). While these also pertain to the QBD $\{(R(t), Z(t), X(t)) : t \geq 0\}$, they are not directly considered by QBD methods since they are implicit in the QBD structure of the generator (transition probability) matrix representation of the system. Nevertheless, the MMPPs become a greater

[1] A probability distribution is *nondefective* if $\lim_{x\to\infty} F(x) = 1$ for a (in this case, continuous) c.d.f $F(\cdot)$. The c.d.f is deemed *proper* if $\int_0^\infty dF(x) = 1$.

focus of emphasis as we must now directly account for their role in determining the stability conditions of the queueing system.

5.2 Derivation of the Semi-Markov Kernel

Define $R_n = R(T_n^+)$, $Z_n = Z(T_n^+)$, and $X_n = X(T_n^+)$ at the instants T_n^+, $n = 0, 1, \ldots, \infty$ just after the server returns to up-and-idle status. The resulting discrete-time process

$$\{((R_n, Z_n, X_n), T_n) : n \geq 0\} \tag{5.1}$$

is then a Markov renewal process with kernel $Q^*(x)$ given by

$$Q^*(x) = \begin{bmatrix} B_0(x) & B_1(x) & B_2(x) & B_3(x) & B_4(x) & \cdots \\ A_0^{(1)}(x) & A_1^{(1)}(x) & A_2^{(1)}(x) & A_3^{(1)}(x) & A_4^{(1)}(x) & \cdots \\ 0 & A_0^{(2)}(x) & A_1^{(2)}(x) & A_2^{(2)}(x) & A_3^{(2)}(x) & \cdots \\ 0 & 0 & A_0^{(3)}(x) & A_1^{(3)}(x) & A_2^{(3)}(x) & \cdots \\ 0 & 0 & 0 & A_0^{(4)}(x) & A_1^{(4)}(x) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

and whose elements, for $i \geq 1$, take the form

$$\begin{aligned} \left[A_\nu^{(i)}(x)\right]_{jj'} = P\{R_{n+1} = i + \nu - 1,\, Z_{n+1} = j',\, \tau_{n+1} \leq x \\ \mid R_n = i,\, Z_{n-1} = j\} \end{aligned} \tag{5.2}$$

and

$$[B_\nu(x)]_{jj'} = P\{R_{n+1} = \nu,\, Z_{n+1} = j',\, \tau_{n+1} \leq x \,|\, R_n = 0,\, Z_n = j\}, \tag{5.3}$$

given $\nu, n \in \mathbb{Z}^+$, $x \in \mathbb{R}^+$, and $i \geq 1$ (see [40: p 157]). Thus, for any regenerative cycle $(T_n, T_{n+1}]$ and for $\nu \geq 0$, the matrices $A_\nu^{(i)}(x)$ contain the probabilities that the orbit size has increased by exactly $\nu - 1$ if the initial orbit size is not zero. The matrix $B_\nu(x)$ contains the probabilities that ν customers arrive during a regenerative cycle that begins with an empty orbit. Because of the fact that the regenerative epochs are the up-and-idle instants, $X_n = 1$ for all $n \geq 0$, and so we will refer to the process (5.1) in the abbreviated form

$$\{((R_n, Z_n), T_n) : n \geq 0\}.$$

It is thus implicit within this expression that $X_n = 1$ for all $n \geq 0$.

We henceforth will refer to the regenerative cycles as *cycles* for the sake of simplicity and define

$$\tau_n \equiv T_{n+1} - T_n, \quad n \geq 0.$$

In order to determine probabilities (5.2) and (5.3), we must first enumerate the possible events that may occur during a cycle. We enumerate these events in a similar manner to that which was done in [62]. When one takes into account the fact that the repair distribution does *not* vary based upon whether the failure was idle or active, we are left with the following scenarios:

1. The server fails before the arrival of a primary customer or a retrial customer (if the orbit is nonempty).

2. A primary customer arrives before a failure or a retrial customer, but the service is interrupted by a failure.

3. A retrial occurs before a failure or a primary customer arrival, but the service is interrupted by a failure.

4. The service of a primary customer is non-interrupted by failure and is completed.

5. The service of a retrial customer is non-interrupted by failure and is completed.

Define $I_n \in \{1, \ldots, 5\}$, $n \in \mathbb{Z}^+$ as the first event that occurs during the cycle $(T_n, T_{n+1}]$. We define the conditional probabilities

$$[\Psi_\nu^{(i)}(k,x)]_{jj'} = \begin{cases} P\{R_{n+1} = i + \nu - 1,\, Z_{n+1} = j', & \text{if } i > 0, \\ \qquad \tau_{n+1} \le x \mid Z_n = j,\, I_n = k,\, R_n = i\} & \\ P\{R_{n+1} = \nu,\, Z_{n+1} = j',\, \tau_{n+1} \le x \mid & \text{if } i = 0, \\ \qquad Z_n = j,\, I_n = k,\, R_n = 0\} & \end{cases}$$

and

$$P_k^{(i)} \;=\; P\left\{I_n = k \,|\, R_n = i\right\}, \qquad i \in \mathbb{Z}^+,\; k = 1, \ldots, 5. \tag{5.4}$$

Since the events $\{1, 2, \ldots, 5\}$ are mutually exclusive, we may utilize the law of total probability to assert that

$$A_\nu^{(i)}(x) \;=\; \sum_{k=1}^{5} \Psi_\nu^{(i)}(k,x) P_k^{(i)}, \tag{5.5}$$

$$B_\nu(x) \;=\; \sum_{k=1}^{5} \Psi_\nu^{(0)}(k,x) P_k^{(i)}, \tag{5.6}$$

It thus becomes clear that, rather than dealing with scalar values, each $A_\nu^{(i)}(x)$ and $B_\nu(x)$ for $i \ge 0$ are now *block* matrices of dimension $m \times m$ (where m denotes the number of states of the Markovian environment).

Our next step is to determine the probabilities $P_k^{(i)}$, $i = 0, 1, 2, \ldots,$ $k = 1, \ldots, 5$. Let $F_\eta(x)$ denote the c.d.f. corresponding to a random variable $X_\eta \sim PH(\boldsymbol{\phi}_\eta, T_\eta)$, where $T_\eta = Q - \Delta(\boldsymbol{\eta})$ (see Appendix, Section A.2). We now proceed to derive (5.4) explicitly for the five events that were defined earlier.

Case $I_n = 1$: The server fails before the arrival of a primary customer or a retrial customer (if the orbit is nonempty).

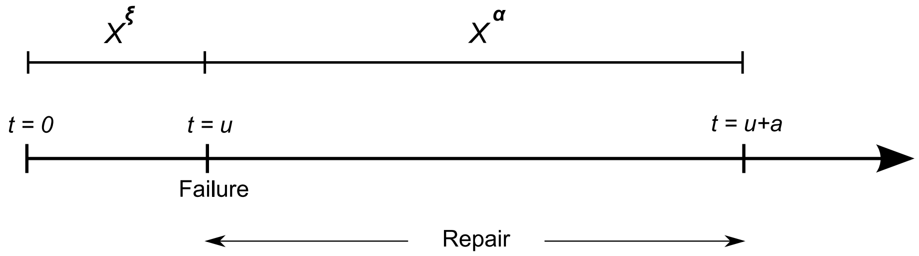

Figure 5.1 Graphical depiction of an idle failure (Case 1).

For this case, we assume the mutual independence of the arrival, uptime, and retrial processes. It is thus necessary to evaluate the probability

$$P_1^{(i)} \equiv P\left\{X_n^{\xi} \leq X_n^{\lambda}\right\} P\left\{X_n^{\xi} \leq X_n^{i\theta}\right\}, \quad i \in \mathbb{Z}^+,$$

where X_n^{ξ} denotes the random uptime of the server, X_n^{λ} gives the interarrival duration, and $X_n^{i\theta}$ is the period of time between retrials, given that there are i customers in the retrial orbit, each of which is measured in the interval $(T_n, T_{n+1}]$. Since $X_n^{\xi} \sim PH(\boldsymbol{\phi}_{\xi}, T_{\xi})$, $X_n^{\lambda} \sim PH(\boldsymbol{\phi}_{\lambda}, T_{\lambda})$, and $X_n^{i\theta} \sim PH(\boldsymbol{\phi}_{i\theta}, T_{i\theta})$, (1.15) gives us

$$P\{X_n^{\xi} \leq X_n^{\lambda}\} = 1 + (\boldsymbol{\phi}_{\xi} \otimes \boldsymbol{\phi}_{\lambda})(-T_{\xi} \oplus T_{\lambda})^{-1}(\boldsymbol{e} \otimes T_{\lambda}\boldsymbol{e}) \tag{5.7}$$

$$P\{X_n^{\xi} \leq X_n^{i\theta}\} = 1 + (\boldsymbol{\phi}_{\xi} \otimes \boldsymbol{\phi}_{i\theta})(-T_{\xi} \oplus T_{i\theta})^{-1}(\boldsymbol{e} \otimes T_{i\theta}\boldsymbol{e}). \tag{5.8}$$

Multiplying (5.7) and (5.8) gives,

$$P_1^{(0)} = 1 + (\boldsymbol{\phi}_{\xi} \otimes \boldsymbol{\phi}_{\lambda})(-T_{\xi} \oplus T_{\lambda})^{-1}(\boldsymbol{e} \otimes T_{\lambda}\boldsymbol{e}), \tag{5.9}$$

$$P_1^{(i)} = P_1^{(0)}[1 + (\boldsymbol{\phi}_{\xi} \otimes \boldsymbol{\phi}_{i\theta})(-T_{\xi} \oplus T_{i\theta})^{-1}(\boldsymbol{e} \otimes T_{i\theta}\boldsymbol{e})], \qquad i \geq 1. \tag{5.10}$$

For arbitrary MMPP rate parameters $\boldsymbol{a}$ and $\boldsymbol{b}$, define

$$P_0(\boldsymbol{a}, \boldsymbol{b}) = P\left\{X_n^{a} \leq X_n^{b}\right\} = 1 + (\boldsymbol{\phi}_{a} \otimes \boldsymbol{\phi}_{b})(-T_{a} \oplus T_{b})^{-1}(\boldsymbol{e} \otimes T_{b}\boldsymbol{e}).$$

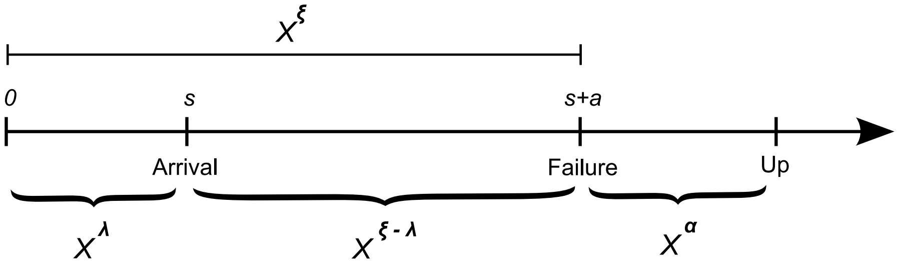

Figure 5.2 Graphical depiction of a failure that occurs during the service of a (primary) customer.

It is clear from the preceding definition that

$$P_0(\boldsymbol{a}, \boldsymbol{b}) \;=\; 1 - P_0(\boldsymbol{b}, \boldsymbol{a}).$$

When applied to equations (5.9) and (5.10), we obtain, respectively,

$$\begin{aligned} P_1^{(0)} &= P_0(\boldsymbol{\xi}, \boldsymbol{\lambda}) \\ P_1^{(i)} &= P_1^{(0)} P_0(\boldsymbol{\xi}, i\boldsymbol{\theta}). \end{aligned}$$

Case $I_n = 2$: A primary customer arrives before a failure or a retrial customer, but the service is interrupted by a failure.

This scenario differs from Case 1 in that we must compute the probability of a failure occurring before the end of a *service* period, which is arbitrarily distributed. Denote the random variable X_n^{μ} to be the total service time *requested* during a cycle $(T_n, T_{n+1}]$ of Type 2. We must state the probability of a failure occurring before the end of a service period in the following manner:

$$P\left\{X_n^{\xi} \le X_n^{\lambda} + X_n^{\mu}\right\} \;=\; P\left\{X_n^{\xi} - X_n^{\lambda} \le X_n^{\mu}\right\}. \tag{5.11}$$

Denote the difference $X_n^{\xi} - X_n^{\lambda}$ by $X_n^{\xi-\lambda}$. We now seek to determine the distribution function $F_{\xi,\lambda}$ associated to the random variable $X_n^{\xi-\lambda}$. To this end, we condition on the arrival time of the primary customer and note, as in Figure 5.2, that if we are given arrival time $s > 0$, then for any $a > 0$, $X_n^{\xi-\lambda} \leq a$ if and only if it is no larger than the interval $[s, s+a]$. This is given by the conditional probability

$$P\left\{X_n^{\xi-\lambda} \leq a \,|\, X_n^{\lambda} \leq X_n^{\xi}\right\}.$$

We must therefore compute the joint probability

$$\begin{aligned} P\{X_n^{\xi-\lambda} \leq a\,,\, X_n^{\lambda} \leq X_n^{\xi}\} \;&=\; \int_0^{\infty}\int_s^{s+a} dF_{\xi}(t)\, dF_{\lambda}(s) \\ &= [\phi_{\xi}(\exp(T_{\xi}a) - I) \otimes \boldsymbol{\phi}_{\lambda}]\, (-T_{\xi} \oplus T_{\lambda})^{-1}\, (e \otimes T_{\lambda}\boldsymbol{e})\,, \end{aligned}$$

from which we obtain

$$\begin{aligned} F_{\xi,\lambda}(a) &= P\{X_n^{\xi-\lambda} \leq a \,|\, X_n^{\lambda} \leq X_n^{\xi}\} \\ &= \frac{P\{X_n^{\xi-\lambda} \leq a\,,\, X_n^{\lambda} \leq X_n^{\xi}\}}{P\{X_n^{\lambda} \leq X_n^{\xi}\}} \\ &= \frac{[\phi_{\xi}(\exp(T_{\xi}a) - I) \otimes \phi_{\lambda}]\, (-T_{\xi} \oplus T_{\lambda})^{-1}\, (e \otimes T_{\lambda}\boldsymbol{e})}{1 + (\boldsymbol{\phi}_{\lambda} \otimes \boldsymbol{\phi}_{\xi})[-T_{\lambda} \oplus T_{\xi}]^{-1}(\boldsymbol{e} \otimes T_{\xi}\boldsymbol{e})}. \end{aligned} \tag{5.12}$$

We next proceed to compute the probabilities of occurrence for Case 2. Once again, invoking the mutual independence of the various MMPPs allows us to state, for $i \geq 0$, that

$$P_2^{(i)} \equiv P\left\{I_n = 2 \,|\, R_n = i\right\} \;=\; P\{X_n^{\lambda} \leq X_n^{\xi}\}\, P\{X_n^{\lambda} \leq X_n^{i\theta}\}\, P\{X_n^{\xi-\lambda} \leq X_n^{\mu}\}. \tag{5.13}$$

For a square matrix M, define the integral operator

$$\Lambda(M) \;=\; \int_0^{\infty} \exp(-Mt)dH(t),$$

where $H(\cdot)$ is the c.d.f. of service time. Using this notation, we expand (5.13) to obtain

$$\begin{aligned}
P_2^{(i)} &= P_0(\boldsymbol{\lambda},\boldsymbol{\xi})P_0(\boldsymbol{\lambda},\boldsymbol{i\theta})\int_0^\infty F_{\xi,\lambda}(u)\,dH(u) \\
&= P_0(\boldsymbol{\lambda},\boldsymbol{i\theta})\left(\boldsymbol{\phi}_\xi\int_0^\infty(\exp(T_\xi u)-I)\;dH(u)\otimes\boldsymbol{\phi}_\lambda\right)(-T_\xi\oplus T_\lambda)^{-1}(\boldsymbol{e}\otimes T_\lambda\boldsymbol{e}) \\
&= P_0(\boldsymbol{\lambda},\boldsymbol{i\theta})\,\big(\boldsymbol{\phi}_\xi(\Lambda(-T_\xi)-I)\otimes\boldsymbol{\phi}_\lambda\big)(-T_\xi\oplus T_\lambda)^{-1}(\boldsymbol{e}\otimes T_\lambda\boldsymbol{e}). \qquad (5.14)
\end{aligned}$$

Expression (5.14) may now be explicitly evaluated if the distribution function $H(\cdot)$ of the service requirements is specified.

Case $I_n = 3$: A retrial occurs before a failure or a primary customer arrival, but the service is interrupted by a failure.

This situation is very similar to that of Case 2, except that the arrival is a retrial customer. If we define the random variable $X^{\xi-i\theta}$ in an analogous manner to $X^{\xi-\lambda}$, then, for $i \geq 1$,

$$\begin{aligned}
P_3^{(i)} &\equiv P\{I_n = 3 \,|\, R_n = i\} \\
&= P\left\{X_n^{i\theta}\leq X_n^\xi\right\}P\left\{X_n^{i\theta}\leq X_n^\lambda\right\}P\left\{X_n^{\xi-i\theta}\leq X_n^\mu\right\} \\
&= P_0(i\boldsymbol{\theta},\boldsymbol{\lambda})\,\left[\boldsymbol{\phi}_\xi\,(\Lambda(-T_\xi)-I)\otimes\boldsymbol{\phi}_{i\theta}\right](-T_\xi\oplus T_{i\theta})^{-1}(\boldsymbol{e}\otimes T_{i\theta}\boldsymbol{e}).
\end{aligned}$$

where $F_{\xi,i\theta}(u)$ is the c.d.f. of the random variable corresponding to the difference $X_n^{\xi-i\theta} = X_n^\xi - X_n^{i\theta}$.

Case $I_n = 4$: The service of a primary customer is non-interrupted by failure and is completed.

Assume now that the end of service occurs before a failure. This may be expressed for $i \geq 0$ as

$$\begin{aligned}
P_4^{(i)} &= P_0(\boldsymbol{\lambda}, \boldsymbol{\xi}) P_0(\boldsymbol{\lambda}, i\boldsymbol{\theta}) \left(1 - P\left\{X_n^{\xi-\lambda} \leq X_n^{\mu}\right\}\right) \\
&= P_0(\boldsymbol{\lambda}, i\boldsymbol{\theta}) \left(P_0(\boldsymbol{\lambda}, \boldsymbol{\xi}) - \boldsymbol{\phi}_\xi \int_0^\infty (\exp(T_\xi u) - I)\, dH(u) \otimes \boldsymbol{\phi}_\lambda \right) \\
&\qquad \times (T_\xi \oplus T_\lambda)^{-1} (\boldsymbol{e} \otimes T_\lambda \boldsymbol{e}) \\
&= P_0(\boldsymbol{\lambda}, i\boldsymbol{\theta}) \left[P_0(\boldsymbol{\lambda}, \boldsymbol{\xi}) - \left(\boldsymbol{\phi}_\xi (\Lambda(-T_\xi) - I) \otimes \boldsymbol{\phi}_\lambda \right) (-T_\xi \oplus T_\lambda)^{-1} (\boldsymbol{e} \otimes T_\lambda \boldsymbol{e}) \right].
\end{aligned}$$

Case $I_n = 5$: The service of a retrial customer runs to completion.

This is the retrial version of Case 4. Accordingly, for $i \geq 1$, we obtain

$$\begin{aligned}
P_5^{(i)} &= P_0(i\boldsymbol{\theta}, \boldsymbol{\xi}) P_0(i\boldsymbol{\theta}, \boldsymbol{\lambda}) \left[1 - P\left\{X_n^{\xi-i\theta} \leq X_n^{\mu}\right\}\right] \\
&= P_0(i\boldsymbol{\theta}, \boldsymbol{\lambda}) \left(P_0(i\boldsymbol{\theta}, \boldsymbol{\xi}) - \boldsymbol{\phi}_\xi \int_0^\infty (\exp(T_\xi u) - I)\, dH(u) \otimes \boldsymbol{\phi}_{i\theta} \right) \\
&\qquad \times (T_\xi \oplus T_{i\theta})^{-1} (\boldsymbol{e} \otimes T_{i\theta} \boldsymbol{e}) \\
&= P_0(i\boldsymbol{\theta}, \boldsymbol{\lambda}) \left[P_0(i\boldsymbol{\theta}, \boldsymbol{\xi}) - \left(\boldsymbol{\phi}_\xi (\Lambda(-T_\xi) - I) \otimes \boldsymbol{\phi}_{i\theta} \right) (-T_\xi \oplus T_{i\theta})^{-1} (\boldsymbol{e} \otimes T_{i\theta} \boldsymbol{e}) \right].
\end{aligned}$$

In much the same case-by-case fashion, we will derive expressions for the generating functions $A^{(i)^*}(z)$ and $B^*(z)$, $i \geq 0$, $|z| \leq 1$, for the block expressions of the semi-Markov kernel $G(x)$. Let us define the Laplace-Stieltjes transform (LST) matrices

$$\widetilde{\Psi}_\nu^{(i)}(k, s) = \int_0^\infty e^{-sx}\, d\Psi_\nu^{(i)}(k, x), \tag{5.15}$$

and their z-transforms with respect to ν as

$$\widetilde{\Psi}^{(i)*}(k, s, z) = \sum_{\nu=0}^\infty \widetilde{\Psi}_\nu^{(i)}(k, s) z^\nu. \tag{5.16}$$

For notational convenience, kernel expressions that do not depend upon the current number in orbit do not carry the superscript i. In such cases, we write

$$\begin{aligned}\Psi_\nu^{(i)}(k,x) &= \Psi_\nu(k,x)\\ \widetilde{\Psi}_\nu^{(i)}(k,s) &= \widetilde{\Psi}_\nu(k,s)\\ \widetilde{\Psi}^{(i)^*}(k,s,z) &= \widetilde{\Psi}^*(k,s,z).\end{aligned}$$

Define the Laplace-Stieltjes transforms of the kernel elements $A_\nu^{(i)}(x)$ and $B_\nu(x)$ of the Markov renewal process $\{((R_n,Z_n),T_n),:n\geq 0\}$ as

$$\begin{aligned}\widetilde{A}_\nu^{(i)}(s) &= \int_0^\infty e^{-sx}\,dA_\nu^{(i)}(x)\\ \widetilde{B}_\nu(s) &= \int_0^\infty e^{-sx}\,dB_\nu(x).\end{aligned}$$

The entries of the transition probability matrix of the Markov chain $\{(R_n,Z_n):n\geq 0\}$ may be obtained by taking the limit of the kernel entries as $x\to\infty$ such that

$$A_\nu^{(i)} = \lim_{x\to\infty} A_\nu^{(i)}(x) = \lim_{s\to 0}\widetilde{A}_\nu^{(i)}(s) \tag{5.17}$$

$$B_\nu = \lim_{x\to 0} B_\nu(x) = \lim_{s\to 0}\widetilde{B}_\nu(s). \tag{5.18}$$

Define the z-transforms

$$\begin{aligned}\widetilde{A}^{(i)*}(s,z) &= \sum_{\nu=0}^\infty z^\nu \widetilde{A}_\nu^{(i)}(s)\\ \widetilde{B}^*(s,z) &= \sum_{\nu=0}^\infty z^\nu \widetilde{B}_\nu(s).\end{aligned}$$

We then form the stochastic matrices $A^{(i)}$ and B as follows:

$$A^{(i)} = \sum_{\nu=0}^{\infty} A_{\nu}^{(i)} = \widetilde{A}^{(i)*}(0,1), \tag{5.19}$$

$$B = \sum_{\nu=0}^{\infty} B_{\nu} = \widetilde{B}^{*}(0,1). \tag{5.20}$$

Now let $\Phi^{\boldsymbol{\eta}}(\nu,t)$ be the MMPP counting process defined in (1.11) of the Appendix. An important fact that we will frequently use is the following, which appears in [40].

Theorem 5.1. *Let $\boldsymbol{\eta}$ be the vector containing the exponential rate parameters of a MMPP $\{Z(t) : t \geq 0\}$ with generator Q. Then*

$$\sum_{\nu=0}^{\infty} \left\{ \int_0^{\infty} \Phi^{\boldsymbol{\eta}}(\nu,t)\, dF_{\eta}(t) \right\} z^{\nu} = \int_0^{\infty} \exp((Q-(1-z)\Delta(\boldsymbol{\eta}))t)\, dF_{\eta}(t). \tag{5.21}$$

Proof. The proof centers around the ability to interchange the integral and sum, which is possible if and only if the function

$$\sum_{\nu=0}^{\infty} \Phi^{\boldsymbol{\eta}}(\nu,t) \tag{5.22}$$

is *uniformly convergent* with respect to $\nu = 0,1,2,\ldots$. Consider the Poisson probability that n arrivals occur in the interval $(0,t]$, which is given by

$$P(n,t) \equiv e^{-\eta t}\frac{(\eta t)^n}{n!}, \qquad n \geq 0, t \geq 0. \tag{5.23}$$

If we let the rate η vary so that $P(n,t) = P(n,t,\eta)$ and take its derivative with respect to η, we obtain

$$\frac{d}{d\eta}P(n,t,\eta) = te^{-\eta t}\frac{(\eta t)^{n-1}}{(n-1)!}\left(1-\frac{\eta t}{n}\right).$$

It is clear that the derivative will be negative for any $\eta > 0$ if we are given large enough t. In other words, $P(n, t, \eta_i)$ will be decreasing for each $i = 1, \dots m$ given $t \geq t^*$ for some $t^* > 0$. Furthermore, if $\eta_i > \eta_j$, then $P(n, t, \eta_i) < P(n, t, \eta_j)$, which then proves that

$$P(n, t, \eta_{\min}) \geq \Phi(n, t), \quad i = 1, \dots m, \tag{5.24}$$

where $\eta_{\min} = \min_{i=1,\dots,m}(\eta_i)$. We also point to the fact that

$$\lim_{t \to \infty} P(n, t, \eta_{\min}) \;=\; 0. \tag{5.25}$$

Due to the verification of conditions (5.24) and (5.25), we may now invoke a well-known theorem of Weierstrass in which these conditions are necessary and sufficient for the uniform convergence of expression (5.22). Consequently, we may now interchange the operations of summation and integration in (5.21) to obtain

$$\begin{aligned}
&\sum_{\nu=0}^{\infty} \left\{ \int_0^{\infty} \Phi^{\boldsymbol{\eta}}(\nu, t)\, dF_\eta(t) \right\} z^\nu \\
&\qquad = \sum_{\nu=0}^{\infty} \left(\int_0^{\infty} P\{N(t) = \nu, Z(t) = j' \,|\, N(0) = 0, Z(0) = j\}\, dF_\eta(t) \right) z^\nu \\
&\qquad = \int_0^{\infty} \left(\sum_{\nu=0}^{\infty} P\{N(t) = \nu, Z(t) = j' \,|\, N(0) = 0, Z(0) = j\}\, z^\nu \right) dF_\eta(t) \\
&\qquad = \int_0^{\infty} \exp((Q - (1 - z)\Delta(\boldsymbol{\eta}))t)\, dF_\eta(t),
\end{aligned}$$

where the final equality follows from Lemma A.2 in the Appendix. □

Since the only entity count that we require from any MMPP is that of customers arriving to the queue, we use the shorthand notation $\Phi(\nu, t) = \Phi^{\boldsymbol{\lambda}}(\nu, t)$ (see (1.11) in the Appendix) for the remainder of the chapter. In addition, define

$$T_\lambda(z) \;=\; Q - (1 - z)\Delta(\boldsymbol{\lambda})\,,$$

which shall serve as a convenient shorthand in the complicated derivations upon which we shall now embark.

It often becomes necessary during the computation of the semi-Markov kernel to count the number of arrivals during repair periods and generally-distributed service times. In order to facilitate this procedure, we define the matrices

$$J_\nu^\eta(x) = \int_0^x \Phi(\nu,t)\,dF_\eta(t), \tag{5.26}$$

$$J_\nu^{\xi,\lambda}(x) = \int_0^x \Phi(\nu,t)\,dF_{\xi,\lambda}(t), \tag{5.27}$$

$$J_\nu^H(x) = \int_0^x \Phi(\nu,t)\,dH(t), \tag{5.28}$$

where $\boldsymbol{\eta}$ is the vector of rate parameters of a MMPP. We will likewise find it necessary to utilize the LST $\widetilde{J}_\nu^\eta(s)$ of $J_\nu^\eta(x)$ and the corresponding z-transform matrix $\widetilde{J}_\nu^{\eta*}(z,s)$, which are defined as

$$\widetilde{J}_\nu^\eta(s) = \int_0^\infty e^{-sx}\,dJ_\nu^\eta(x), \tag{5.29}$$

$$\widetilde{J}^{\eta^*}(s,z) = \sum_{\nu=0}^\infty z^\nu \int_0^\infty e^{-sx}\,dJ_\nu^\eta(x). \tag{5.30}$$

The following theorem, which is a restatement of [82: Thm 5.1.5], allows the explicit computation of (5.29) evaluated at $s=0$:

Theorem 5.2. *The matrix*

$$\widetilde{J}_\nu^\eta(0) = \int_0^\infty \Phi(\nu,x)\,dF_\eta(x)$$

is given by

$$\widetilde{J}_0^\eta(0) = (I\otimes\boldsymbol{\phi}_\eta)(T_\lambda\oplus T_\eta)^{-1}(I\otimes T_\eta\boldsymbol{e}) \tag{5.31}$$

$$\widetilde{J}^{\eta}_{\nu}(0) \;=\; DC^{\nu-1}E, \qquad \nu \geq 1$$

where the matrices D, C, and E are given by

$$\begin{aligned}
D &= (I \otimes \boldsymbol{\phi}_{\eta})(T_{\lambda} \oplus T_{\eta})^{-1}(T_{\lambda}\boldsymbol{e} \otimes I) \\
C &= (\boldsymbol{\phi}_{\lambda} \otimes I)(T_{\lambda} \oplus T_{\eta})^{-1}(T_{\lambda}\boldsymbol{e} \otimes I) \\
E &= (\boldsymbol{\phi}_{\lambda} \otimes I)(T_{\lambda} \oplus T_{\eta})^{-1}(I \otimes T_{\eta}\boldsymbol{e}).
\end{aligned}$$

The transform matrix (5.30) may, in turn, be evaluated at $s = 0$ and $z = 1$ using Theorem 5.1:

$$\begin{aligned}
\widetilde{J}^{\eta^*}(0,1) &= (I \otimes \phi_{\eta}) \sum_{\nu=0}^{\infty} z^{\nu} \int_0^{\infty} \Phi(\nu, x) \otimes \exp(T_{\eta}x)\, dx (I \otimes T_{\eta}\boldsymbol{e}) \bigg|_{z=1} \\
&= (I \otimes \phi_{\eta}) \int_0^{\infty} \sum_{\nu=0}^{\infty} z^{\nu} \Phi(\nu, x) \otimes \exp(T_{\eta}x)\, dx (I \otimes T_{\eta}\boldsymbol{e}) \bigg|_{z=1} \\
&= (I \otimes \phi_{\eta}) \int_0^{\infty} \exp(Q - (1-z)\Delta(\boldsymbol{\lambda})\, x) \otimes \exp(T_{\eta}x)\, dx (I \otimes T_{\eta}\boldsymbol{e}) \bigg|_{z=1} \\
&= (I \otimes \phi_{\eta}) \int_0^{\infty} \exp\left([Q \oplus T_{\eta}]x\right)\, dx (I \otimes T_{\eta}\boldsymbol{e}) \\
&= (I \otimes \phi_{\eta}) \left(-Q \oplus T_{\eta}\right)^{-1} (I \otimes T_{\eta}\boldsymbol{e}).
\end{aligned}$$

We will now complete the derivation of the semi-Markov kernel block entries $A^{(i)}_{\nu}$ and B_{ν} as defined in (5.17).

Case $I_n = 1$: The server fails before the arrival of a primary customer or a retrial customer (if the orbit is nonempty).

In this scenario the up-idle period ends with a failure, which is then followed by a repair. No customers are processed during this cycle, and thus

$$\Psi_0(2,x) \;=\; 0 \quad \text{for } x \geq 0.$$

For $\nu \geq 1$, the cycle commences with an up-idle period followed by a failure, and then a repair period during which primary arrivals may occur. Hence, we employ the renewal argument by conditioning upon the time of failure as follows:

$$\begin{aligned}\Psi_\nu(1,x) &= \int_0^x J^\alpha_{\nu-1}(x-t)\,dF_\xi(t),\\ &= \int_0^x \int_0^{x-t} \Phi(\nu-1,u)\,dF_\alpha(u)\,dF_\xi(t),\end{aligned} \tag{5.32}$$

where $J^\alpha_\nu(x)$ is the matrix defined in (5.26). Since we require the derivative of (5.32), it is necessary to reduce this expression to one that requires only a single integration. We thus employ the technique of partial integration to obtain the equivalence

$$\begin{aligned}\Psi_\nu(1,x) \;=\; J^\alpha_{\nu-1}(x) - (I\otimes\phi_\xi\otimes\phi_\alpha)\int_0^x (\Phi(\nu-1,t)\otimes\exp(((-T_\xi)\oplus T_\alpha)t))\,dt\\ \times\,(I\otimes\exp(T_\xi x)\boldsymbol{e}\otimes T_\alpha\boldsymbol{e}).\end{aligned} \tag{5.33}$$

We next turn our attention to the evaluation of the LST of $\Psi_\nu(1,x)$, $\widetilde{\Psi}_\nu(1,s)$, which may be obtained as follows:

$$\begin{aligned}\widetilde{\Psi}_\nu(1,s) \;=\; &\int_0^\infty e^{-sx}\,d\Psi_\nu(1,x)\\ =\; &\widetilde{J}^\alpha_{\nu-1}(s) - (I\otimes\phi_\xi\otimes\phi_\alpha)\Big\{\int_0^\infty \big(e^{-sx}\Phi(\nu-1,x)\otimes I\otimes\exp(T_\alpha x)T_\alpha\boldsymbol{e}\big)\,dx\\ &+\int_0^\infty e^{-sx}\int_0^x \Phi(\nu,t)\otimes\exp(((-T_\xi)\oplus T_\alpha)t)\,dt\end{aligned}$$

$$\times \left(I \otimes \exp(T_\xi x) T_\xi \boldsymbol{e} \otimes T_\alpha \boldsymbol{e}\right) dx \Big\}$$

$$= \widetilde{J}^{\alpha}_{\nu-1}(s), \tag{5.34}$$

from which we may compute the conditional probabilities $\Psi_\nu(1,\infty)$ of the occurrence of $\nu - 1$ arrivals to the retrial orbit in a cycle in which $I_n = 1$:

$$\Psi_\nu(1,\infty) \equiv \lim_{x\to\infty} \Psi_\nu(1,x) \;=\; \widetilde{\Psi}_\nu(1,s)\Big|_{s=0} \;=\; \widetilde{J}^{\alpha}_{\nu-1}(0). \tag{5.35}$$

It is possible to obtain a closed-form solution of (5.35) through the application of Theorem 5.2. In this way, we determine that

$$\Psi_\nu(1,\infty) \;=\; \widetilde{J}^{\alpha}_{\nu}(0) \;=\; \begin{cases} (I \otimes \boldsymbol{\phi}_\alpha)(T_\lambda \oplus T_\alpha)^{-1}(I \otimes T_\alpha \boldsymbol{e}) & \text{if } \nu = 0 \\ \\ DC^{\nu-1}E & \text{if } \nu \geq 1 \end{cases}, \tag{5.36}$$

where the matrices D, C, and E are given by

$$\begin{aligned} D &= (I \otimes \boldsymbol{\phi}_\alpha)(T_\lambda \oplus T_\alpha)^{-1}(T_\lambda \boldsymbol{e} \otimes I) \\ C &= (\boldsymbol{\phi}_\lambda \otimes I)(T_\lambda \oplus T_\alpha)^{-1}(T_\lambda \boldsymbol{e} \otimes I) \\ E &= (\boldsymbol{\phi}_\lambda \otimes I)(T_\lambda \oplus T_\alpha)^{-1}(I \otimes T_\alpha \boldsymbol{e}). \end{aligned}$$

Our final task is to derive the generating function

$$\widetilde{\Psi}^*(1,s,z) = \sum_{\nu=0}^{\infty} \widetilde{\Psi}_\nu(1,s)\, z^\nu.$$

From (5.34) we thus obtain

$$\begin{aligned}\widetilde{\Psi}^*(1,0,z) &= z^0 \cdot 0 + \sum_{\nu=1}^{\infty} z^\nu \widetilde{J}^\alpha_{\nu-1}(0,s) \\ &= z\widetilde{J}^{\alpha^*}(0,s,z) \\ &= (I \otimes \phi_\alpha) z \left(-T_\lambda(z) \oplus T_\alpha\right)^{-1} (I \otimes T_\alpha \boldsymbol{e}).\end{aligned}$$

Evaluating z at 1 gives

$$(I \otimes \phi_\alpha) \left(-Q \oplus T_\alpha\right)^{-1} (I \otimes T_\alpha \boldsymbol{e}). \tag{5.37}$$

The quantity (5.37), as well as its counterparts $\widetilde{\Psi}^*(k,0,1)$, $k = 2,\ldots,5$, will be required to derive the stochastic matrices (5.19) and (5.20).

Case $I_n = 2$: A primary customer arrives before a failure or a retrial customer, but the service is interrupted by a failure.

We count arrivals during both of an interrupted service epoch and the repair period that follows. Because the customer in service will inevitably join the retrial orbit, the definition of $\Psi_\nu(2,x)$ implies that

$$\Psi_\nu(2,x) = 0 \quad \text{for } \nu = 0,1.$$

For $\nu \geq 2$, it is necessary to consider arrivals over both of a truncated service and repair epochs, the sum total of which must equal $\nu - 2$. Let $X^{\xi-\lambda+\alpha}$ denote the random variable for the interval of time during which such an interrupted service and repair takes place. As in Case 1, we condition upon the arrival time t of the first customer arrival to obtain, for $t \in (0, x)$ and $\nu \geq 2$, the expression

$$\begin{aligned}\Psi_\nu(2, x) &= \int_0^x J_{\nu-2}^{\xi,\lambda,\alpha}(x-t)\, dF_\lambda(t) \qquad (5.38)\\ &= \int_0^x \int_0^{x-t} \Phi(\nu-2, u)\, dF_{\xi,\lambda,\alpha}(u)\, dF_\lambda(t),\end{aligned}$$

where we denote the c.d.f. of the distribution of the combined interrupted service and repair intervals by $F_{\xi,\lambda,\alpha}(t)$ and define

$$J_\nu^{\xi,\lambda,\alpha}(x) = \int_0^x \Phi(\nu, u)\, dF_{\xi,\lambda,\alpha}(u).$$

In an analogous manner to Case 1, we arrive at the following expression for the probability of the orbit size incrementing by $\nu - 1$ during a cycle in which $I_n = 2$:

$$\Psi_\nu(2, \infty) = \widetilde{\Psi}_\nu(0) = \widetilde{J}_{\nu-2}^{\xi,\lambda,\alpha}(0) = \int_0^\infty \Phi(\nu-2, x)\, dF_{\xi,\lambda,\alpha}(x).$$

In order to evaluate this integral any further, we must determine the c.d.f. $F_{\xi,\lambda,\alpha}(t)$ of the sum $X^{\xi-\lambda+\alpha} = X^{\xi-\lambda} + X^\alpha$. This can be computed in the following manner:

$$\begin{aligned}F_{\xi,\lambda,\alpha}(x) &= \int_0^x F_\alpha(x-t) dF_{\xi,\lambda}(t)\\ &= F_{\xi,\lambda}(x) - (\boldsymbol{\phi}_\alpha \otimes \boldsymbol{\phi}_\xi \otimes \boldsymbol{\phi}_\alpha) \int_0^x \exp(((-T_\alpha) \oplus T_\xi)t) \otimes I\, dt\\ &\qquad \times \left[\exp(T_\alpha x)\boldsymbol{e} \otimes (-T_\xi \oplus T_\lambda)^{-1}(e \otimes T_\lambda \boldsymbol{e})\right]\\ &= F_{\xi,\lambda}(x) - (\boldsymbol{\phi}_\alpha \otimes \boldsymbol{\phi}_\xi \otimes \boldsymbol{\phi}_\alpha)\\ &\qquad \times \left\{\left[(\exp(((-T_\alpha) \oplus T_\xi)x) - I \otimes I)\right] \otimes I\right\}\\ &\qquad \times \left[\exp(T_\alpha x)\boldsymbol{e} \otimes (-T_\xi \oplus T_\lambda)^{-1}(e \otimes T_\lambda \boldsymbol{e})\right]. \qquad (5.39)\end{aligned}$$

From here, we proceed directly to the computation of the z-transform $\widetilde{\Psi}^{(i)*}(2,0,z)$ evaluated at $z=1$. As in Case 1, we begin with the definition of the z-transform to obtain

$$\begin{aligned}
\widetilde{\Psi}^{(i)*}(2,0,z) \;&=\; \left[0\cdot z^0 + 0\cdot z + \sum_{\nu=2}^{\infty} z^{\nu}\,\widetilde{J}^{\xi,\lambda,\alpha}_{\nu-2}(0)\right] \;=\; z^2\,\widetilde{J}^{\xi,\lambda,\alpha^*}(0,z) \\
&= \Big\{(I\otimes\boldsymbol{\phi}_\xi\otimes\boldsymbol{\phi}_\lambda)\left[z^2(-T_\lambda(z)\oplus T_\xi)^{-1}\otimes I\right]\left[I\otimes(T_\xi\otimes I)\right]\Big\}/P_0(\boldsymbol{\lambda},\boldsymbol{\xi}) \\
&\quad - (I\otimes\boldsymbol{\phi}_\alpha\otimes\boldsymbol{\phi}_\xi\otimes\boldsymbol{\phi}_\alpha)\Big\{\left[z^2(-T_\lambda(z)\otimes I\oplus T_\xi)^{-1}\otimes I\right] \\
&\quad \times \big\{(I\otimes(-T_\alpha)\boldsymbol{e}\otimes(I\otimes I)) + (I\otimes\boldsymbol{e}\otimes(T_\xi\otimes I)) + (I\otimes T_\alpha\boldsymbol{e}\otimes(I\otimes I))\big\} \\
&\quad - \left[z^2(-T_\lambda(z)\oplus T_\alpha)^{-1}\otimes(I\otimes I)\right](I\otimes T_\alpha\boldsymbol{e}\otimes(I\otimes I))\Big\}(-T_\xi\oplus T_\lambda)^{-1}(e\otimes T_\lambda e).
\end{aligned}$$

Case $I_n = 3$: A retrial occurs before a failure or a primary customer arrival, but the service is interrupted by a failure.

We now assume that a failure occurs during the service of a retrial customer, and thus the current orbit size i does not increment by one at the failure epoch. Therefore, $\Psi^{(i)}_0(3,x)=0$ and, for $\nu\geq 1$,

$$\begin{aligned}
\Psi^{(i)}_{\nu}(3,x) \;&=\; \int_0^x J^{\xi,i\theta,\alpha}_{\nu-1}(x-t)\,dF_{i\theta}(t) && (5.40)\\
&=\; \int_0^x\int_0^{x-t}\Phi(\nu-1,u)\,dF_{\xi,i\theta,\alpha}(u)\,dF_{i\theta}(t),
\end{aligned}$$

where we denote the c.d.f. of the distribution of the interrupted service and repair interval by $F_{\xi,i\theta,\alpha}(t)$ and define

$$J^{\xi,i\theta,\alpha}_{\nu}(x) \;=\; \int_0^x \Phi(\nu,u)\,dF_{\xi,i\theta,\alpha}(u).$$

With this established, it is a simple matter to adapt the results of Case 2 in making following statements, the first being

$$\Psi_\nu(3,\infty) \;=\; \widetilde{\Psi}_\nu(3,0) \;=\; \widetilde{J}^{\xi,i\theta,\alpha}_{\nu-1}(0) \;=\; \int_0^\infty \Phi(\nu-1,x)\,dF_{\xi,i\theta,\alpha}(x),$$

where the c.d.f. $F_{\xi,i\theta,\alpha}(x)$ of the interrupted service and repair distribution is given by

$$\begin{aligned} F_{\xi,i\theta,\alpha}(x) &= \int_0^x F_\alpha(x-t)dF_{\xi,i\theta}(t) \\ &= F_{\xi,i\theta}(x) - (\boldsymbol{\phi}_\alpha \otimes \boldsymbol{\phi}_\xi \otimes \boldsymbol{\phi}_\alpha) \\ &\qquad \times \Big\{ \big[(\exp(((-T_\alpha) \oplus T_\xi)x) - I \otimes I)\big] \otimes I \Big\} \\ &\qquad \times \big[\exp(T_\alpha x)\boldsymbol{e} \otimes (-T_\xi \oplus T_{i\theta})^{-1}(e \otimes T_{i\theta}\boldsymbol{e})\big]. \end{aligned} \tag{5.41}$$

The z-transform expressions $\widetilde{\Psi}^{(i)*}(3,0,z)$ may likewise be computed as

$$\begin{aligned} \widetilde{\Psi}^{(i)*}(3,0,z) &= \left[0 \cdot z^0 + \sum_{\nu=1}^{\infty} z^\nu \widetilde{J}^{\xi,i\theta,\alpha}_{\nu-1}(0)\right] = z\widetilde{J}^{\xi,i\theta,\alpha^*}(0,z) \\ &= \Big\{(I \otimes \boldsymbol{\phi}_\xi \otimes \boldsymbol{\phi}_{i\theta})\left[z(-T_{i\theta}(z) \oplus T_\xi)^{-1} \otimes I\right]\left[I \otimes (T_\xi \otimes I)\right]\Big\}/P_0(\boldsymbol{i\theta},\boldsymbol{\xi}) \\ &\quad - (I \otimes \boldsymbol{\phi}_\alpha \otimes \boldsymbol{\phi}_\xi \otimes \boldsymbol{\phi}_\alpha)\Big\{\left[z(-T_{i\theta}(z) \otimes I \oplus T_\xi)^{-1} \otimes I\right] \\ &\quad \times \big\{(I \otimes (-T_\alpha)\boldsymbol{e} \otimes (I \otimes I)) + (I \otimes \boldsymbol{e} \otimes (T_\xi \otimes I)) + (I \otimes T_\alpha\boldsymbol{e} \otimes (I \otimes I))\big\} \\ &\quad - \left[z(-T_{i\theta}(z) \oplus T_\alpha)^{-1} \otimes (I \otimes I)\right](I \otimes T_\alpha\boldsymbol{e} \otimes (I \otimes I))\Big\}(-T_\xi \oplus T_{i\theta})^{-1}(e \otimes T_{i\theta}e). \end{aligned}$$

Case $I_n = 4$: The service of a primary customer is non-interrupted by failure and is completed.

Conditioning on the first arrival time, we obtain the following for $\nu \geq 1$, where $J_{\nu}^{H}(\cdot)$ is defined as in (5.28):

$$\Psi_{\nu}(4,x) \;=\; \int_{0}^{x} J_{\nu-1}^{H}(x-t)dF_{\lambda}(t) \;=\; \int_{0}^{x}\int_{0}^{x-t} \Phi(\nu-1,u)\,dH(u)dF_{\lambda}(t). \qquad (5.42)$$

As before, we use partial integration to establish the equivalence

$$\begin{aligned}
&\int_{0}^{x}\int_{0}^{x-t} \Phi(\nu-1,u)\,dH(u)dF_{\lambda}(t) \\
&\qquad = \int_{0}^{x} \Phi(\nu-1,x-t)F_{\lambda}(t)\,dH(x-t) \;=\; \int_{0}^{x} \Phi(\nu-1,u)F_{\lambda}(x-u)\,dH(u) \\
&\qquad = J_{\nu-1}^{H}(x) - (I\otimes\phi_{\lambda})\int_{0}^{x} \Phi(\nu-1,u)\otimes\exp(-T_{\lambda}u)\,dH(u)\,(I\otimes\exp(T_{\lambda}x)).
\end{aligned}$$

The LST of this expression may be computed as

$$\begin{aligned}
\widetilde{\Psi}_{\nu}(4,s)\Big|_{s=0} \;=\;& \widetilde{J}_{\nu-1}^{H}(0) - (I\otimes\boldsymbol{\phi}_{\lambda})\Big[(I\otimes\exp(T_{\lambda}x)\boldsymbol{e}\,dH(x) \\
&+\int_{0}^{\infty}\int_{0}^{x}(\Phi(\nu-1,u)\otimes\exp(-T_{\lambda}u))(I\otimes\exp(T_{\lambda}x)T_{\lambda}\boldsymbol{e})\,dH(u)\Big],
\end{aligned}$$

from which we obtain

$$\widetilde{\Psi}_{\nu}(4,0) \;=\; \widetilde{J}_{\nu-1}^{H}(0).$$

The z-transform $\widetilde{\Psi}^{*}(4,0,z)$ is thus given by

$$\begin{aligned}
\widetilde{\Psi}^{*}(4,0,z) \;&=\; z^{0}\cdot 0 + z\sum_{\nu=1}^{\infty} z^{\nu-1}\left[\int_{0}^{\infty}\Phi(\nu-1,x)\,dH(x)\right] \\
&=\; z\int_{0}^{\infty}\exp(T_{\lambda}(z)x)\,dH(x) \\
&=\; z\Lambda(-T_{\lambda}(z)),
\end{aligned}$$

which gives

$$\widetilde{\Psi}^{*}(4,0,1) \;=\; \Lambda(-Q). \tag{5.43}$$

Case $I_n = 5$: The service of a retrial customer is non-interrupted by failure and is completed.

This case is the only one of the five in which the orbit size may decrement. The appropriate integral expression for $\nu \geq 0$ is thus given by

$$\Psi^{(i)}_{\nu}(5,x) \;=\; \int_0^x J^H_{\nu}(x-t)dF_{i\theta}(t) \;=\; \int_0^x \int_0^{x-t} \Phi(\nu,u)\, dH(u)dF_{i\theta}(t). \tag{5.44}$$

From Case 4, we thus have that

$$\widetilde{\Psi}_{\nu}(5,0) \;=\; \widetilde{J}^H_{\nu}(0).$$

The corresponding probability generating function is therefore given by

$$\widetilde{\Psi}^{(i)^*}(5,0,z) \;=\; \Lambda(-T_{\lambda}(z)),$$

from whence we conclude that

$$\widetilde{\Psi}^{(i)^*}(5,0,1) \;=\; \Lambda(-Q). \tag{5.45}$$

With the kernel of the embedded chain $\{(R_n, Z_n) : n \geq 0\}$ now specified, we may investigate the conditions for the ergodicity of the system discussed in this chapter. An additional prerequisite to the determination of stability conditions is the establishment of stability criteria for level-dependent discrete-time $M/G/1$-type Markov chains using the theory of ergodicity for general Markov chains. This will enable us to determine the exact criteria for the stability of the embedded chain, whose recurrence properties mirror that of the process in which it is embedded.

5.3 Stability Analysis

In this section, we shall present criteria for the positive recurrence (and hence, ergodicity) of a general level-dependent $M/G/1$-type Markov chain $\{(L_n, Y_n) : n \geq 0\}$. Following this, we will derive an explicit stability formula for the embedded process $Q^*(\infty)$ of the Markov-modulated unreliable $M/G/1$ retrial queueing system. We first provide the necessary definitions, which are the level-dependent analogues to those provided in Section 3.2.4. The determination of stability criteria for a $M/G/1$-type Markov chain revolves around the stochastic matrix $A^{(i)}$, which we previously defined in (5.19):

$$A^{(i)} = \sum_{\nu=0}^{\infty} A_\nu^{(i)}, \tag{5.46}$$

where $A_\nu^{(i)} = A_\nu^{(i)}(\infty)$ and $i \geq 0$. In order to apply what was done for level-independent discrete-time $M/G/1$-type Markov chains in Section 3.2.4, we define the following for $i \geq 1$:

$$\beta^{(i)} = \sum_{\nu=1}^{\infty} \nu A_\nu^{(i)}. \tag{5.47}$$

and

$$\rho^{(i)} = \boldsymbol{\pi}^{(i)} \beta^{(i)} \boldsymbol{e}, \tag{5.48}$$

where we define $\boldsymbol{\pi}^{(i)}$ to be the invariant probability vector of the stochastic matrix $A^{(i)}$. Consequently, $\boldsymbol{\pi}^{(i)}$ contains as entries the steady-state probabilities of $\{Y_n : n \geq 0\}$, given a starting orbit size of $i \geq 1$.

In order to apply the ergodicity results of [87] and [95], we must, nevertheless, determine the level-dependent drift $d(i)$ of the embedded chain $\{(L_n, Y_n) : n \geq 0\}$. This may be obtained by noting that, from definition (5.5) of $A_\nu^{(i)}$, where $i = 1, 2, 3, \ldots$ and $\nu = 0, 1, 2, \ldots$, the increment of the process is *not* ν, but is actually

$\nu - 1$, given that there are $i \geq 1$ in the retrial orbit at the beginning of a cycle. In light of this observation, we arrive at the following Lemma:

Lemma 5.1. *Suppose that the discrete-time process* $\{(L_n, Y_n) : n \geq 0\}$ *is a* $M/G/1$*-type Markov chain. Then its drift,* $d(i)$ *is given by*

$$d(i) \;=\; \rho^{(i)} - 1.$$

Proof. We proceed using the realizations $\nu - 1$, where $\nu = 0, 1, 2, \ldots$, of the random increments of $\{L_n : n \geq 0\}$ and the definition of expectation:

$$\begin{aligned} d(i) &= \boldsymbol{\pi}^{(i)} \sum_{\nu=0}^{\infty} (\nu - 1) A_\nu^{(i)} \boldsymbol{e} \\ &= \boldsymbol{\pi}^{(i)} \sum_{\nu=0}^{\infty} \nu A_\nu^{(i)} - \boldsymbol{\pi}^{(i)} \sum_{\nu=0}^{\infty} A_\nu^{(i)} \boldsymbol{e} \\ &= \boldsymbol{\pi}^{(i)} \beta^{(i)} \boldsymbol{e} - \boldsymbol{\pi}^{(i)} A^{(i)} \boldsymbol{e} \\ &= \rho^{(i)} - 1. \end{aligned} \tag{5.49}$$

□

Thus, the quantity (5.49) is the *drift at level* i for an $M/G/1$-type Markov chain. It is clear that $\rho^{(i)} \leq 1$ holds if, and only if, $d(i) \leq 0$, and so we designate $\rho^{(i)}$ as the *conditional traffic intensity formula for the embedded Markov chain* $\{(L_n, Y_n) : n \geq 0\}$.

As with any DTMC, the recurrence properties of the $M/G/1$-type process are predicated on the behavior of the drift (5.49) across levels of the process. We use this relationship to define a simple condition that is equivalent to the ergodicity of an (aperiodic) $M/G/1$-type DTMC $\{(L_n, Y_n) : n \geq 0\}$, where L_n denotes the *level* of the process. It is clear from Pakes' Lemma that the necessary condition for positive recurrence of $\{(L_n, Y_n) : n \geq 0\}$ is that $\limsup_{i \to \infty} d(i) < 0$, or, equivalently,

$\limsup_{i\to\infty} \rho^{(i)} < 1$. We now specify this condition for the limit, rather than the limit supremum, in order to demonstrate a sufficient condition for positive recurrence. It is once more necessary to utilize Theorem 3.7 in validating this criterion, which, in turn, requires that we demonstrate Kaplan's condition (see Section 3.1.3) for the process $\{(L_n, Y_n)\}$.

Showing that Kaplan's condition holds is equivalent to proving that the sequence of real-valued functions $\{\psi_i(z) : i \in \mathbb{Z}^+,\ z \in [0,1)\}$ (see (3.25)) is bounded from below. This sequence, written explicitly, becomes

$$\begin{aligned}
\psi_i(z) &= \left\{ z^i - \sum_{\nu=0}^{\infty} p_\nu^{(i)} z^{i'} \right\} / (1-z) \\
&= \left\{ z^i - \sum_{\nu=0}^{\infty} (\boldsymbol{\pi}^{(i)} A_\nu^{(i)} \boldsymbol{e}) z^{i+\nu-1} \right\} / (1-z) \\
&= \left\{ z^i - \boldsymbol{\pi}^{(i)} \sum_{\nu=0}^{\infty} A_\nu^{(i)} z^{i+\nu-1} \boldsymbol{e} \right\} / (1-z), \qquad (5.50)
\end{aligned}$$

where $p_\nu^{(i)}$ is the probability that the size of the orbit transitions from i to $i+\nu-1$. It is necessary to show that there exist numbers $c \in (0,1)$, $B > 0$, and $N > 0$ such that $\psi_i(z) \geq -B$ for $z \in [c, 1)$ and $i \geq N$. In other words, there must exist real numbers $c \in (0,1)$ and $N > 0$ such that

$$\inf\{\psi_i(z) : z \in [c,1)\} \geq -B,$$

holds for every $i \geq N$. Such a determination is complicated by the presence of the infinite sum in (5.50), and so a calculus approach to the minimization of $\psi_i(z)$ would be a formidable task. We therefore choose an alternate route to the verification of Kaplan's condition for the $M/G/1$-type DTMC, one that focuses upon the conditions listed in Theorem 3.8.

Theorem 5.3. *An irreducible, aperiodic, and discrete-time level-dependent $M/G/1$-type Markov chain $\{(L_n, Y_n) : n \geq 0\}$ is positive recurrent if and only if the following conditions hold:*

$$\lim_{i\to\infty} \rho^{(i)} < 1, \qquad \textit{and} \qquad \sum_{\nu=0}^{\infty} \nu B_\nu < \infty, \tag{5.51}$$

where $\rho^{(i)}$ is defined as in (5.48).

Proof. The outline of the proof is generally the same as that for Theorem 4.2, except that we now investigate the stability of the Markov chain $\{(L_n, Y_n) : n \geq 0\}$. We begin with the assumption that (1) $\rho^{(i)} < 1$, and (2) $\sum_{\nu=0}^{\infty} \nu B_\nu < \infty$. It is clear from (5.49) that (1) implies that the drift $d(i)$ is negative in the limit. In order to facilitate the discussion of (2), we observe that

$$d(0) \;=\; \sum_{\nu=0}^{\infty} \nu B_\nu < \infty.$$

This, combined with the decreasing nature of the drift over levels of the process, demonstrates that the remaining drift terms for $i \geq 1$ are likewise finite. Invoking Pakes' lemma as before, we thus obtain the positive recurrence of the Markov chain.

We next prove the reverse implication, namely that in which we assume the hypothesis does *not* hold and $\{(L_n, Y_n)\}$ is ergodic. We shall derive a contradiction using [95: Thm 1], which requires the fulfillment of Kaplan's condition as an initial step. As previously mentioned, we will avoid an explicit verification of this condition, and, instead, proceed to verify condition (2) of Theorem 3.8. This requires a proof of the existence of a lower bound for the sequence $\{\delta_i : i = 0, 1, 2, \ldots\}$, where

$$\delta_i \;=\; \sum_{j \leq i} p_{ij}(j - i),$$

and p_{ij} is the (i, j)th element of the transition probability matrix P of $\{(L_n, Y_n)\}$. We refine the expression for δ_i by noting that, for the state space $\mathcal{S}$ of the process

$\{(L_n, Y_n)\}$,

$$\min\{j - i : i \geq j, i, j \in \mathcal{S}\} = -1,$$

since the only decrement (i.e., transition to a lower level) that a $M/G/1$-type process is allowed is -1. Applying this observation leads to the following for $i \geq 1$:

$$\begin{aligned}
\delta_i &= \sum_{j \leq i} p_{ij}(j - i) \\
&= (-\boldsymbol{\pi}^{(i-1)} A_0^{(i-1)} + 0 \cdot \boldsymbol{\pi}^{(i)} A_1^{(i)})\boldsymbol{e} \\
&= -\boldsymbol{\pi}^{(i-1)} A_0^{(i-1)} \boldsymbol{e} \\
&\geq -1
\end{aligned}$$

with (5.52) a result of the fact that $\boldsymbol{\pi}^{(i)} A_\nu^{(i)} \boldsymbol{e}$ is the unconditional probability of an increment of $\nu - 1$ in the level for $\nu \geq 0$. Hence, we have shown that Kaplan's condition holds.

Finally, we recall that from Lemma 5.1 that the drift at level i is given by $d(i) = \rho^{(i)} - 1$. Since we assume that $\lim_{i\to\infty} \rho^{(i)} > 1$, there must exist some $N > 0$ such that $d(i) > 0$ whenever $i \geq N$. By Theorem 3.7, we conclude that the $M/G/1$-type DTMC $\{(L_n, Y_n)\}$ cannot be ergodic, which contradicts our assumption. This result completes the proof of necessity. □

We now return to the discussion of the stability of the embedded chain $Q^*(\infty)$, which we have shown to be a Markov chain of $M/G/1$-type. In order to facilitate the derivation of simple analytic criteria for $Q^*(\infty)$ to be positive recurrent, we define the (limiting) matrices

$$\bar{A} = \lim_{i\to\infty} A^{(i)} \text{ and } \beta = \lim_{i\to\infty} \beta^{(i)} = \lim_{i\to\infty} \sum_{\nu=0}^{\infty} \nu A_\nu^{(i)},$$

and the vector

$$\boldsymbol{\pi} = \lim_{i\to\infty} \boldsymbol{\pi}^{(i)},$$

assuming that the limits exist. It is then true by the fundamental properties of limits that

$$\rho = \lim_{i\to\infty} \boldsymbol{\pi}^{(i)}\beta^{(i)}\boldsymbol{e} = \boldsymbol{\pi}\beta\boldsymbol{e},$$

or, in other words, we may first compute individual limits, then multiply the terms together in order to obtain ρ. Thus, it remains to determine the methods by which we compute the matrices $\beta^{(i)}$ and the invariant probability vectors $\boldsymbol{\pi}^{(i)}$.

The determination of $\beta^{(i)}$ is contingent on making the observation that, since the matrix z-transform, $A^{(i)^*}(z) = \sum_{\nu=0}^{\infty} A_\nu^{(i)} z^\nu$, is analytic on the unit disk $|z| \leq 1$, we may interchange the order of the derivative and the summation to obtain

$$\frac{d}{dz}A^{(i)^*}(z) = \sum_{\nu=0}^{\infty} \frac{d}{dz}\left(A_\nu^{(i)} z^\nu\right) = \sum_{\nu=0}^{\infty} \nu A_\nu^{(i)} z^{\nu-1}, \tag{5.52}$$

and thus we may obtain the matrix $\beta^{(i)}$ by taking the derivative of the matrix-transform $A^{(i)^*}(z)$ and evaluating at $z = 1$, as follows:

$$\beta^{(i)} = \left.\frac{d}{dz}A^{(i)^*}(z)\right|_{z=1}. \tag{5.53}$$

The vector $\boldsymbol{\pi} = \lim_{i\to\infty} \boldsymbol{\pi}^{(i)}$, where the $\boldsymbol{\pi}^{(i)}$ are the invariant vectors of the matrices $A^{(i)}$, is obtained by noting the following, which is that

$$\boldsymbol{\pi} = \lim_{i\to\infty} \boldsymbol{\pi}^{(i)} = \lim_{i\to\infty} \boldsymbol{\pi}^{(i)}A^{(i)} = \boldsymbol{\pi}\bar{A} \tag{5.54}$$

and

$$1 = \lim_{i\to\infty} \boldsymbol{\pi}^{(i)}\boldsymbol{e} = \boldsymbol{\pi}\boldsymbol{e}, \tag{5.55}$$

and thus $\boldsymbol{\pi}$ is the invariant probability vector of the stochastic matrix $\bar{A}$. The computation of $\bar{A}$ itself may be accomplished via the following:

$$A^{(i)} = \sum_{\nu=0}^{\infty} A_\nu^{(i)} = \sum_{\nu=0}^{\infty} A_\nu^{(i)} z^\nu \Bigg|_{z=1} = A^{(i)^*}(1)$$

(see (5.46) for the definition of $A^*(z)$). We thus obtain

$$\bar{A} = \lim_{i\to\infty} A^{(i)^*}(1). \tag{5.56}$$

5.3.1 Computation of the Traffic Intensity

With the theoretical background afforded by Theorem 5.3 we may now proceed to determine the values of the parameter vectors $\boldsymbol{\lambda}$, $\boldsymbol{\xi}$, $\boldsymbol{\alpha}$, $\boldsymbol{\theta}$, and the scalar-valued average service time $1/\mu$ that guarantee the validity of the following statements:

$$\lim_{i\to\infty} \boldsymbol{\pi}^{(i)} \beta^{(i)} \boldsymbol{e} \;=\; \lim_{i\to\infty} \boldsymbol{\pi}^{(i)} \sum_{\nu=0}^{\infty} \nu A_\nu^{(i)} \boldsymbol{e} \;<\; 1, \tag{5.57}$$

$$\sum_{\nu=0}^{\infty} \nu B_\nu \;<\; \infty. \tag{5.58}$$

Through unconditioning, and by the definition of $\widetilde{\Psi}_\nu^{(i)}(k,0)$, we obtain

$$\widetilde{A}_\nu^{(i)} \;=\; \sum_{k=1}^{5} \widetilde{\Psi}_\nu^{(i)}(k,0) P_k^{(i)}. \tag{5.59}$$

Define

$$\widetilde{M}_k^{(i)} \;=\; \frac{d}{dz} \widetilde{\Psi}^{(i)*}(k,s,z) \Bigg|_{\substack{s=0\\ z=1}},$$

which is equivalent to the expected number of primary arrivals to the queue in a (regenerative) cycle given that there are i customers in the system (orbit). If one factors in the observation that

$$\beta^{(i)} = \sum_{\nu=0}^{\infty} \nu A_{\nu}^{(i)} = \frac{d}{dz}\widetilde{A}^{(i)*}(s,z)\Big|_{\substack{s=0\\z=1}}, \tag{5.60}$$

$$\sum_{\nu=0}^{\infty} \nu B_{\nu} = \frac{d}{dz}\widetilde{B}^{*}(s,z)\Big|_{\substack{s=0\\z=1}}, \tag{5.61}$$

it becomes clear that we may rewrite (5.57) and (5.58) as

$$\lim_{i\to\infty} \boldsymbol{\pi}^{(i)}\beta^{(i)}\boldsymbol{e} = \lim_{i\to\infty} \boldsymbol{\pi}^{(i)} \left\{ \sum_{k=1}^{5} \widetilde{M}_k^{(i)} P_k^{(i)} \right\} \boldsymbol{e}, \tag{5.62}$$

and

$$\sum_{\nu=0}^{\infty} \nu B_{\nu} = \sum_{k=1}^{5} \widetilde{M}_k^{(0)} P_k^{(0)} < \infty. \tag{5.63}$$

Starting from (5.56), we obtain the invariant probability vector of the stochastic matrix $\bar{A}$ as follows:

$$\bar{A} = A^{(i)*}(1) = \lim_{i\to\infty} \sum_{\nu=0}^{\infty} z^{\nu} \widetilde{A}_{\nu}^{(i)}(s)\Big|_{\substack{s=0\\z=1}}.$$

From (5.59), we thus obtain

$$\bar{A} = \lim_{i\to\infty} \sum_{\nu=0}^{\infty} z^{\nu} \sum_{k=1}^{5} \widetilde{\Psi}_{\nu}^{(i)}(k,s) P_k^{(i)}\Big|_{\substack{s=0\\z=1}} = \sum_{k=1}^{5} \widetilde{\Psi}^{(i)*}(k,0,1) P_k^{(i)}. \tag{5.64}$$

The interchange of summations in (5.64) is contingent upon the analyticity of $\widetilde{\Psi}_{\nu}^{(i)}(k,s)$ in a neighborhood of $s = 0$.

5.3.2 Limiting Values of $P_k^{(i)}$

We begin by considering the terms $P_k = \lim_{i\to\infty} P_k^{(i)}$ for cycle types $k = 1, 2, \ldots, 5$. To this end, we present the following result:

Proposition 5.1.

$$\lim_{i\to\infty} (T_\eta \oplus T_{i\theta})^{-1} \;=\; \lim_{i\to\infty} (T_{i\theta} \oplus T_\eta)^{-1} \;=\; \mathbf{0},$$

where $T_\eta = Q - \Delta(\boldsymbol{\eta})$ for some exponential rate parameter $\boldsymbol{\eta}$.

Proof. We recall from Lemma A.1 that

$$\begin{aligned}(T_\eta \oplus T_{i\theta})^{-1} &= \int_0^\infty \exp\left[(-T_\eta \oplus T_{i\theta})x\right]\, dx \\ &= \int_0^\infty \exp(-T_\eta x) \otimes \exp(T_{i\theta} x)\, dx \\ &= \int_0^\infty \exp(-T_\eta x) \otimes \exp\left[(Q - i\Delta(\boldsymbol{\theta}))x\right]\, dx.\end{aligned}$$

since Q and $\Delta(\boldsymbol{\theta})$ commute, we may further simplify the above to obtain

$$(T_\eta \oplus T_{i\theta})^{-1} \;=\; \int_0^\infty \exp(-T_\eta x) \otimes \exp(Qx) \exp\left[-i\Delta(\boldsymbol{\theta})\, x\right]\, dx. \tag{5.65}$$

Since we have, for diagonal matrices $\Delta(\boldsymbol{\eta})$, the identity

$$\exp\left(\Delta(\boldsymbol{\eta})\right) \;=\; \Delta\left(\exp(\boldsymbol{\eta})\right)$$

where $\exp(\boldsymbol{\eta}) = [e^{\eta_j}]_{j\in\{1,\ldots,m\}}$, (5.65) becomes

$$\int_0^\infty \exp(-T_\eta x) \otimes \exp(Qx) \Delta\left(e^{-i\boldsymbol{\theta}x}\right)\, dx \to \mathbf{0}$$

as $i \to \infty$, and thus $\lim_{i\to\infty} (T_\eta \oplus T_{i\theta})^{-1} \;=\; \mathbf{0}$. □

Proposition 5.1 allows us to prove the following for a generic vector $\boldsymbol{a}$:

Proposition 5.2. *Given a (constant) generic MMPP rate vector $\boldsymbol{a}$ and*

$$P_0(i\boldsymbol{\theta}, \boldsymbol{a}) \;=\; P\left\{X^{i\theta} \leq X^a\right\},$$

its limit as $i \to \infty$ *is given by*

$$\lim_{i\to\infty} P_0(i\boldsymbol{\theta}, \boldsymbol{a}) = \lim_{i\to\infty} [1 + (\boldsymbol{\phi}_{i\theta} \otimes \boldsymbol{\phi}_a)(-T_{i\theta} \oplus T_a)^{-1}(\boldsymbol{e} \otimes (T_a\boldsymbol{e})] \;=\; 1. \tag{5.66}$$

Proof. The result immediately follows from the application of Proposition 5.1. □

The next, and last, result resolves an issue concerning the limiting value of a certain type of expression containing MMPP-related elements:

Proposition 5.3. *Given a (constant) generic MMPP rate vector* $\boldsymbol{a}$ *and a generic vector* $\boldsymbol{\gamma}$*, we obtain the limiting value*

$$\begin{aligned}
&\lim_{i\to\infty} (\boldsymbol{\phi}_{i\theta} \otimes \boldsymbol{\gamma})(-T_{i\theta} \oplus T_a)^{-1}(T_{i\theta}\boldsymbol{e} \otimes \boldsymbol{e}) \qquad (5.67)\\
&\quad = \lim_{i\to\infty} (\boldsymbol{\phi}_\gamma \otimes \boldsymbol{\phi}_{i\theta})(-T_a \oplus T_{i\theta})^{-1}(\boldsymbol{e} \otimes T_{i\theta}\boldsymbol{e})\\
&\quad = -\boldsymbol{\gamma}\boldsymbol{e}.
\end{aligned}$$

Proof. As in Proposition 5.1, we rewrite the argument of the limit in (5.67) in integral form and then apply the technique of partial integration:

$$\begin{aligned}
&(\boldsymbol{\phi}_{i\theta} \otimes \boldsymbol{\gamma})(T_{i\theta} \oplus T_a)^{-1}(T_{i\theta}\boldsymbol{e} \otimes \boldsymbol{e})\\
&\qquad = \int_0^\infty (\boldsymbol{\phi}_{i\theta} \exp(-T_{i\theta}t)T_{i\theta}\boldsymbol{e})(\boldsymbol{\gamma} \exp(T_a t)\boldsymbol{e})\, dt\\
&\qquad = (\boldsymbol{\phi}_{i\theta} \exp(-T_{i\theta}t)\boldsymbol{e})(\boldsymbol{\gamma} \exp(T_a t)\boldsymbol{e})\Big|_{t=0}^{\infty}\\
&\qquad\qquad\qquad - (\boldsymbol{\phi}_{i\theta} \otimes \boldsymbol{\gamma}) \int_0^\infty \exp((T_{i\theta} \oplus T_\xi)t)\, dt(\boldsymbol{e} \otimes T_a\boldsymbol{e})\\
&\qquad = -(\boldsymbol{\phi}_{i\theta}\boldsymbol{e})(\boldsymbol{\gamma}\boldsymbol{e}) - 0 \;=\; -\boldsymbol{\gamma}\boldsymbol{e}.
\end{aligned}$$

□

Through the application of Proposition 5.1, one may conclude that

$$P_k = \lim_{i\to\infty} P_k^{(i)} = 0, \quad k = 1, 2, 4, \tag{5.68}$$

since each are multiplied by a probability of the form $P_0(\boldsymbol{a}, i\boldsymbol{\theta})$, where $\boldsymbol{a}$ is a generic MMPP rate vector. For Cases 3 and 5, Proposition 5.3 may be used to show that

$$\begin{aligned} P_3 \equiv \lim_{i\to\infty} P_3^{(i)} &= \boldsymbol{\phi}_\xi \left(I - \Lambda(-T_\xi)\right) \boldsymbol{e} \\ P_5 \equiv \lim_{i\to\infty} P_5^{(i)} &= 1 - \boldsymbol{\phi}_\xi \left(I - \Lambda(-T_\xi)\right) \boldsymbol{e}. \end{aligned} \tag{5.69}$$

This result confirms the intuitive notion that Case 3, which is the interruption of the service of a retrial customer, and Case 5, the completion of a retrial service, are the key events that influence system stability as the orbit size becomes large. Further revealing is the observation that the scenario of Case 5 is the only one that permits the orbit size to decrement; thus, at a minimum, the probabilities of transition between orbit sizes resulting from the situation of Case 5 must be taken into consideration for all $i \geq 1$.

5.3.3 The Ergodicity of $\{(R_n, Z_n) : n \geq 0\}$

Now that the requisite facts and concepts have been presented, we may compute an explicit traffic intensity formula for the embedded chain $\{(R_n, Z_n) : n \geq 0\}$. We shall first state the result, and then provide the actual steps of the computation as a proof.

Theorem 5.4. *The embedded Markov chain* $\{(R_n, Z_n) : n \geq 0\}$ *is ergodic if and only if both of the following conditions hold:*

1. *The overall traffic intensity,* ρ, *fulfills the condition*

$$\rho = \boldsymbol{p}\beta\boldsymbol{e} < 1,$$

where $\boldsymbol{p}$ *is the steady-state probability vector of the random environment,* β *is the matrix given by*

$$\beta = \Big\{\Upsilon_3(-Q) + \Upsilon_3(\Delta(\boldsymbol{\lambda}))\Big\}P_3 + \left\{\frac{d}{dz}\Lambda(-T_\lambda(z))\bigg|_{z=1}\right\}P_5, \tag{5.70}$$

and

$$\begin{aligned}\Upsilon_3(X) &= (I \otimes \boldsymbol{\phi}_\xi)\left[(Q \oplus T_\xi)^{-1}(X \oplus T_\xi)(Q \oplus T_\xi)^{-1}\right](I \otimes T_\xi \boldsymbol{e}) \\ &\quad - (I \otimes \boldsymbol{\phi}_\alpha \otimes \boldsymbol{\phi}_\xi)\Big\{\left[(Q \otimes I \oplus T_\xi)^{-1}(X \otimes I \oplus T_\xi)(Q \otimes I \oplus T_\xi)^{-1}\right] \\ &\quad \times \left[(I \otimes (-T_\alpha)\boldsymbol{e} \otimes \boldsymbol{e}) + (I \otimes \boldsymbol{e} \otimes T_\xi \boldsymbol{e})\right] \\ &\quad - \left[(Q \oplus T_\alpha)^{-1}(X \oplus T_\alpha)(Q \oplus T_\alpha)^{-1} \otimes I)\right](I \otimes T_\alpha \boldsymbol{e} \otimes \boldsymbol{e})\Big\}.\end{aligned} \tag{5.71}$$

2. *The condition,*

$$\begin{aligned}&(I \otimes \boldsymbol{\phi}_\alpha)(Q \oplus T_\alpha)^{-1}\Big\{(I \otimes T_\alpha \boldsymbol{e}) + (\Delta(\boldsymbol{\lambda}) \oplus T_\alpha)(Q \oplus T_\alpha)^{-1}\Big\} \\ &\quad \times (I \otimes T_\alpha \boldsymbol{e})\, P_1^{(0)} + \Big\{2\Upsilon_2(-Q) + \Upsilon_2(\Delta(\boldsymbol{\lambda}))\Big\}P_2^{(0)} \\ &\quad + \left\{\Lambda(-Q) + \frac{d}{dz}\Lambda(-T_\lambda(z))\bigg|_{z=1}\right\} P_5^{(0)} < \infty,\end{aligned} \tag{5.72}$$

where

$$\begin{aligned}\Upsilon_2(X) &= \Big\{(I \otimes \boldsymbol{\phi}_\xi \otimes \boldsymbol{\phi}_\lambda)\left[(Q \oplus T_\xi)^{-1}(X \oplus T_\xi)(Q \oplus T_\xi)^{-1} \otimes I\right] \\ &\quad \times \left[I \otimes (T_\xi \otimes I)\right]\Big\}/P_0(\boldsymbol{\lambda}, \boldsymbol{\xi}) - (I \otimes \boldsymbol{\phi}_\alpha \otimes \boldsymbol{\phi}_\xi \otimes \boldsymbol{\phi}_\alpha)\end{aligned}$$

$$\times \Bigg\{ \big[(Q \otimes I \oplus T_\xi)^{-1}(X \otimes I \oplus T_\xi)(Q \otimes I \oplus T_\xi)^{-1} \otimes I\big]$$

$$\times \big\{(I \otimes (-T_\alpha)\boldsymbol{e} \otimes (I \otimes I)) + (I \otimes \boldsymbol{e} \otimes (T_\xi \otimes I)) + (I \otimes T_\alpha \boldsymbol{e} \otimes (I \otimes I))\big\}$$

$$- \big[(Q \oplus T_\alpha)^{-1}(X \oplus T_\alpha)(Q \oplus T_\alpha)^{-1} \otimes (I \otimes I)\big]$$

$$\times (I \otimes T_\alpha \boldsymbol{e} \otimes (I \otimes I))\Bigg\}(-T_\xi \oplus T_\lambda)^{-1}(e \otimes T_\lambda e). \quad (5.73)$$

Proof. This result is a straightforward application of Theorem 5.3, which provides the basic criteria for the ergodicity of any $M/G/1$-type DTMC. The first stipulation of the theorem is that

$$\boldsymbol{\pi}\beta\boldsymbol{e} < 1,$$

where $\boldsymbol{\pi}$ is defined as the invariant probability vector of

$$\bar{A} = \lim_{i\to\infty} \sum_{\nu=0}^{\infty} A_\nu^{(i)}.$$

It is clear from this definition that $\bar{A}$ gives the steady-state transition probabilities of $\{(R_n, Z_n) : n \geq 0\}$ aggregated by environment state, which is the same regardless of the number i in orbit. Thus, $\boldsymbol{\pi} = \boldsymbol{p}$, which is interpreted as the steady-state probability vector of the random environment embedded at regenerative epochs.

We next use (5.62) as a basis for the computation of ρ in condition (1), while applying the results for the limiting probabilities of Cases 1 through 5 given in (5.68) and (5.69) in order to state

$$\begin{aligned} \beta &= \lim_{i\to\infty} \beta^{(i)} = \lim_{i\to\infty} \sum_{k=1}^{5} \widetilde{M}_k^{(i)}\, P_k^{(i)} \\ &= \lim_{i\to\infty} \left(\widetilde{M}_3^{(i)}\, P_3^{(i)} + \widetilde{M}_5^{(i)}\, P_5^{(i)}\right) \end{aligned}$$

$$= \lim_{i\to\infty} \frac{d}{dz} \left(\widetilde{\Psi}^{(i)*}(3,s,z)\, P_3^{(i)} + \widetilde{\Psi}^{(i)*}(5,s,z)\, P_5^{(i)} \right) \Bigg|_{\substack{s=0\\z=1}} . \tag{5.74}$$

Tedious algebra then yields the expressions given in (5.70) and (5.71).

Subsequently, for condition (2), we refer to the second criterion of Theorem 5.3, which is that

$$\sum_{\nu=0}^{\infty} \nu B_\nu < \infty.$$

As in (5.61), we compute this quantity in the following manner:

$$\begin{aligned}
\sum_{\nu=0}^{\infty} \nu B_\nu &= \sum_{k=1}^{5} \widetilde{M}_k^{(0)}\, P_k^{(0)} \\
&= \sum_{k=1}^{5} \frac{d}{dz} \widetilde{\Psi}^{(0)*}(k,s,z) P_k^{(0)} \Bigg|_{\substack{s=0\\z=1}} \\
&= \frac{d}{dz} \left(\widetilde{\Psi}^{(0)*}(3,s,z) P_3^{(0)} + \widetilde{\Psi}^{(0)*}(5,s,z) P_5^{(0)} \right) \Bigg|_{\substack{s=0\\z=1}} .
\end{aligned}$$

This likewise yields the quantities (5.72) and (5.73), which estabilishes (2) as the remaining criterion for the ergodicity of $\{(R_n, Z_n)\}$. We have thus shown that conditions (1) and (2) are the necessary and sufficient requirements for the ergodicity of $\{(R_n, Z_n)\}$ as specified by Theorem 5.3. □

As with the traffic intensity formula (4.18), all terms containing θ disappear in the limit. This is in accordance with results for all other retrial queues, which is that the retrial rate has no bearing upon the long-term stability of the queue. We may use Theorem 5.4 to interpret this result, which is, plainly stated, that system stability is measured against the *worst* possible scenario, namely that in which the orbit size grows very large. However, as the orbit size grows large, the retrial rates also grow without bound, until the system becomes an instantaneous-feedback retrial queue (i.e., from the point of view of the server, retrials occur continuously). It is thus apparent that a retrial will obtain service with probability one, *regardless of the value of the individual retrial rate.*

5.3.4 Application to the Exponential Model

Theorem 5.4 applies to the embedded Markov chain $\{(R_n, Z_n) : n \geq 0\}$ of $\{(R(t), Z(t), X(t)) : t \geq 0\}$ at the regenerative epochs T_n so long as the service distribution meets the requirements stated in Section 5.1. We may thus compare the stability criteria for the general model to that with exponential service. More specifically, we consider the retrial queueing model of [62] that excludes the possibilities of customer balking and Markov modulation, with the additional assumption of exponentially-distributed service times. This model is equivalent to $\{(R(t), Z(t), X(t)) : t \geq 0\}$ if one fixes all of the exponential rates of arrival, service, failure, repair, and retrial. Any random environment that is an irreducible Markov chain may be thus chosen for the comparison.

The parameters of the model itself are not numerically specified. They are simply left in terms of their symbolic representations λ, μ, ξ, α, and θ, as defined in Sections 4.1 and 5.1. The density function $h(x)$ of the service distribution is thus chosen to be

$$h(x) = \mu e^{-\mu x},$$

which is the exponential p.d.f. with rate μ. We define the infinitesimal generator Q of the random environment as the 2×2 matrix

$$Q = \begin{bmatrix} -1 & 1 \\ 2 & -2 \end{bmatrix}$$

whose steady-state probability vector is thus given by $\boldsymbol{p} = [2/3, 1/3]$. Lastly, all of the initial probabilities $\boldsymbol{\phi}_\eta$, where $\boldsymbol{\eta}$ is the rate vector of a MMPP, are fixed at $\boldsymbol{p}$.

The next task is to obtain the limiting probabilities for Cases 3 and 5. Direct computation yields

$$P_3 = \frac{\xi}{\mu+\xi} \quad \text{and} \quad P_5 = \frac{\mu}{\mu+\xi}.$$

The corresponding 2×2 transform matrices

$$\lim_{i\to\infty} \widetilde{\Psi}^{(i)*}(k,0,z), \quad k = 3,5,$$

are found to be

$$\widetilde{\Psi}^{(i)*}(3,0,z) = \frac{z\alpha}{(\lambda(1-z)+\alpha)(\lambda(1-z)+\alpha+3)} \begin{bmatrix} \lambda(1-z)+\alpha+2 & 1 \\ 2 & \lambda(1-z)+\alpha+1 \end{bmatrix},$$

and

$$\widetilde{\Psi}^{(i)*}(5,0,z) = \frac{1}{\mu+3} \begin{bmatrix} \mu+2 & 1 \\ 2 & \mu+1 \end{bmatrix}.$$

In order to compute the matrix β, it is sufficient to obtain

$$\lim_{i\to\infty} \widetilde{M}^{(i)*}(3,0,1) = \frac{d}{dz} \lim_{i\to\infty} \widetilde{\Psi}^{(i)*}(3,0,z)\bigg|_{z=1},$$

since $\widetilde{\Psi}^{(i)*}(5,0,z)$ is constant. Consequently,

$$\beta = \lim_{i\to\infty} \widetilde{M}^{(i)*}(3,0,1)P_3^{(i)} = \frac{d}{dz} \lim_{i\to\infty} \widetilde{\Psi}^{(i)*}(3,0,z)P_3^{(i)}\bigg|_{z=1}$$

$$= \frac{\lambda\mu\xi}{(\mu+\xi)(\lambda(1-z)+\mu)^2\lambda(1-z)+\mu+3)^2}$$

$$\times \begin{bmatrix} k_1(z) & (2\lambda(1-z)+2\mu+3)^2 \\ -(2\lambda(1-z)+2\mu+3)^2 & k_2(z) \end{bmatrix}\Bigg|_{z=1},$$

where

$$k_1(z) = \frac{2}{3}(\lambda(1-z)+\mu+3)^2 + \frac{1}{3}(\lambda(1-z)+\mu)^2,$$

$$k_2(z) = \frac{1}{3}(\lambda(1-z)+\mu+3)^2 + \frac{2}{3}(\lambda(1-z)+\mu)^2.$$

The traffic intensity expression may finally be computed as

$$\rho \, = \, \boldsymbol{p}\beta\boldsymbol{e} \, = \, \frac{\lambda\xi + \lambda(\alpha+\xi)}{\alpha(\mu+\xi)}. \tag{5.75}$$

Notice that (5.75) is exactly the conditional traffic intensity formula that can be derived from [62: Thm 1], and which appears in (4.34).

This concludes the discussion of the steady-state analysis of the unreliable $M/G/1$ retrial queue in a random environment. We were able, using the matrix-analytic theory of Markov chains of $M/G/1$-type, to obtain the transition probability matrix for embedded Markov chain of a complex retrial queueing system. We subsequently proved a condition for the ergodicity of *general*, level-dependent, discrete-time $M/G/1$-type Markov chains. To the best of our knowledge these results are novel. We then utilized this condition in order to derive a stability expression for the unreliable Markov-modulated $M/G/1$ retrial queueing system.

6. Queueing Model Optimization

The results of Chapter 4 may be used to improve the performance of retrial queueing systems through the optimization of design or operating parameters. However, the matrix-analytic approach taken in Chapter 4 leads to a numerical approximation of the steady-state orbit size distribution; therefore, a closed-form objective function of the operating parameters is not available. We therefore resort to mesh-adaptive search techniques for problems that have no derivative information. For such techniques, it is not necessary to specify a closed-form objective function, so long as the objective function may be numerically evaluated at points inside the feasible region. For detailed information on the search algorithms applied in this chapter, the reader is referred to the excellent summaries contained in [2, 15, 16]. This chapter is illustrative in nature, highlighting the potential usefulness of the main results (for the exponential model) in improving the performance of retrial queues with complex dynamics.

6.1 Problem Formulation

A principal concern of most queueing systems is the average amount of time customers spend in the retrial orbit. For instance, retrial queues can be used to model customer contact centers wherein customers who dial into the center may not receive service immediately and try back after a random time. A critical performance measure is the expected time such customers spend outside the system attempting to gain service. An intrinsic cost (that may be difficult to ascertain) may be associated to each unit of time that the customer spends in the orbit. Thus, we consider the minimization of $E[W_R]$. The decision variables are the exponential arrival and service rates defined by the vectors $\boldsymbol{\lambda} = [\lambda_j]$ and $\boldsymbol{\mu} = [\mu_j]$, respectively. We denote by $\rho(\boldsymbol{\lambda}, \boldsymbol{\mu})$, the traffic intensity defined in (4.18), now expressed as a function of the decision variables $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$.

As before, the integer m denotes the number of distinct random environment states, and we define the nonnegative row vectors $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{y}_1$, and $\boldsymbol{y}_2$ such that for $j = 1, 2, \ldots, m$,

$$\begin{aligned}
\boldsymbol{x}_1 &= [x_{11}, x_{12}, \ldots, x_{1j}], \\
\boldsymbol{x}_2 &= [x_{21}, x_{22}, \ldots, x_{2j}], \\
\boldsymbol{y}_1 &= [y_{11}, y_{12}, \ldots, y_{1j}], \\
\boldsymbol{y}_2 &= [y_{21}, x_{22}, \ldots, y_{2j}]
\end{aligned}$$

with $\boldsymbol{x}_1 < \boldsymbol{x}_2$ and $\boldsymbol{y}_1 < \boldsymbol{y}_2$. We may then state the mathematical programming formulation as follows:

$$\begin{aligned}
\min \quad & E[W_R] \\
\text{s.t.} \quad & \rho(\boldsymbol{\lambda}, \boldsymbol{\mu}) < 1 && \text{(6.1a)} \\
& \lambda_j \in [x_{1j}, x_{2j}], \quad j = 1, \ldots, m && \text{(6.1b)} \\
& \mu_j \in [y_{1j}, y_{2j}], \quad j = 1, \ldots, m. && \text{(6.1c)}
\end{aligned}$$

It should be recognized that the objective function $E[W_R]$ may only be evaluated numerically, while constraints (6.1b) and (6.1c) are box constraints that are provided by the user. The only quantity that may be computed explicitly from the model is the left-hand side of constraint (6.1a), which is the traffic intensity (4.18).

6.2 Solution Procedure

The optimization software NOMADm [1] was employed to solve problem formulation (6.1). The advantage of this approach rests in the intrinsic flexibility of this software, which can locate high-quality solutions given only a numerically defined objective function, a set of constraints and an initial feasible solution. NOMADm employs a form of *generalized pattern search* (GPS) algorithm known as *mesh-adaptive* *direct search* (MADS), that incrementally explores, or *polls*, the surrounding feasible region at each iteration of the procedure. The iterations of the algorithm are conducted on a *mesh*, which is a means by which the increments of steps through the feasible region is controlled. The mesh may be coarsened or refined as the situation dictates, hence the term *mesh-adaptive.* The first step is entitled *search*; any point on the mesh may be evaluated during a search routine, which, as stated in [2], can be defined in any manner whatsoever. The *poll* step, however, is essential to the workings of the algorithm. It involves the search of points on the mesh that neighbor the current iterate. Most of the theoretical focus regarding generalized pattern search concerns the effect of various polling strategies upon rate of convergence of the algorithm and the quality of the solutions.

Because MADS is stochastic in the way that it polls surrounding points, it is instructive to conduct multiple optimization runs for the same problem instance. The user may alleviate the computational cost of such additional runs by storing previously visited points in a cache file, which thus allows faster run times, though at the expense of hard-disk storage. Selecting different initial feasible points for the algorithm may also increase confidence that a global rather than local minimum has been obtained. A suitably chosen search routine, in conjunction with polling, may likewise be used to mitigate this risk.

It should be noted that, in the numerical implementation of (6.1), constraint (6.1a) is not included as a formal constraint. Rather, any points $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ that render the system unstable are simply not investigated any further during a polling (or search) step. The algorithm removes such points from its list of solution candidates and then proceeds to the next trial point on the mesh. This is done as a precaution to ensure that the routine does not attempt to evaluate the objective function at points that lead to instability of the queueing system. In such cases, the steady-state probabilities do not exist, and hence, we avoid further problems that may arise from this circumstance.

6.3 Numerical Illustrations

The problem instances presented in this subsection are exponential with 3- and 5-state random environments, respectively. Initial feasible vectors $\boldsymbol{\lambda}_0$ and $\boldsymbol{\mu}_0$ were specified for each of the 3- and 5-state cases. The remaining vectors, $\boldsymbol{\xi}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$, were assigned values that did not change during optimization runs. Whenever a run terminated, the values of $\boldsymbol{\lambda}^*$ and $\boldsymbol{\mu}^*$ that produced the optimal cost $E[W_R]^*$ were recorded. Additional runs, using distinct initial feasible points, were used to help ensure that the algorithm produced consistent solutions. As noted above, the objective is to minimize the steady-state mean time spent in orbit, $E[W_R]$.

6.3.1 Three-State Environment

We now present results for the NOMADm optimization of the queueing model subject to a three-state random environment with infinitesimal generator matrix

$$Q = \begin{bmatrix} -2.0 & 1.0 & 1.0 \\ 1.0 & -2.0 & 1.0 \\ 1.0 & 1.0 & -2.0 \end{bmatrix}. \tag{6.2}$$

The values of the parameters $\boldsymbol{\xi}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\theta}$ were fixed at the values shown in Table 6.1. In addition, we specify the box constraints on the decisions variables $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ to be

$$1 \quad \leq \lambda_j \leq 6, \quad j = 1, 2, 3 \tag{6.3}$$

$$1 \quad \leq \mu_j \leq 8, \quad j = 1, 2, 3. \tag{6.4}$$

Results for a single optimization run appear in Table 6.2. These intuitive results indicate that the optimal parameters $\boldsymbol{\lambda}^*$ and $\boldsymbol{\mu}^*$ lie mostly on the boundaries of the region defined by (6.3) and (6.4). This is a consequence of the fact that no

Table 6.1 Problem data for 3-state example.

Vector	State 1	State 2	State 3
$\boldsymbol{\lambda}_0$	2.0	3.5	4.0
$\boldsymbol{\mu}_0$	4.0	3.5	7.0
$\boldsymbol{\xi}$	3.0	2.0	0.1
$\boldsymbol{\alpha}$	4.0	7.0	2.0
$\boldsymbol{\theta}$	1.5	3.0	5.0

penalties were assessed in the problem formulation for increasing service rates or for decreasing customer traffic. An exception to this was the value of λ_3^*, which lies in the interior of the region defined by the box constraints (6.3) and (6.4). That $\lambda_3^* \geq 1$ was not binding is due to the intersection of the region enclosed by the box constraints with that defined by the constraint $\rho(\boldsymbol{\lambda}, \boldsymbol{\mu}) < 1$.

We were likewise interested in the convergence rate of MADS. A plot of the iteration history (see Figure 6.1) revealed that most of the optimality gap is eliminated by MADS in the first 50 of 389 iterations. This information is needed for estimating truncation points for individual runs, particularly as the number of environmental states increase. We shall obtain a glimpse of the dimensionality problem in the five-state example that we discuss next, for which computation time becomes a significant issue.

Table 6.2 Optimal solution for 3-state example.

Vector	State 1	State 2	State 3	$E[W_R]^*$
$\boldsymbol{\lambda}^*$	1.00	1.00	1.71	6.9014
$\boldsymbol{\mu}^*$	8.0	8.0	8.0	

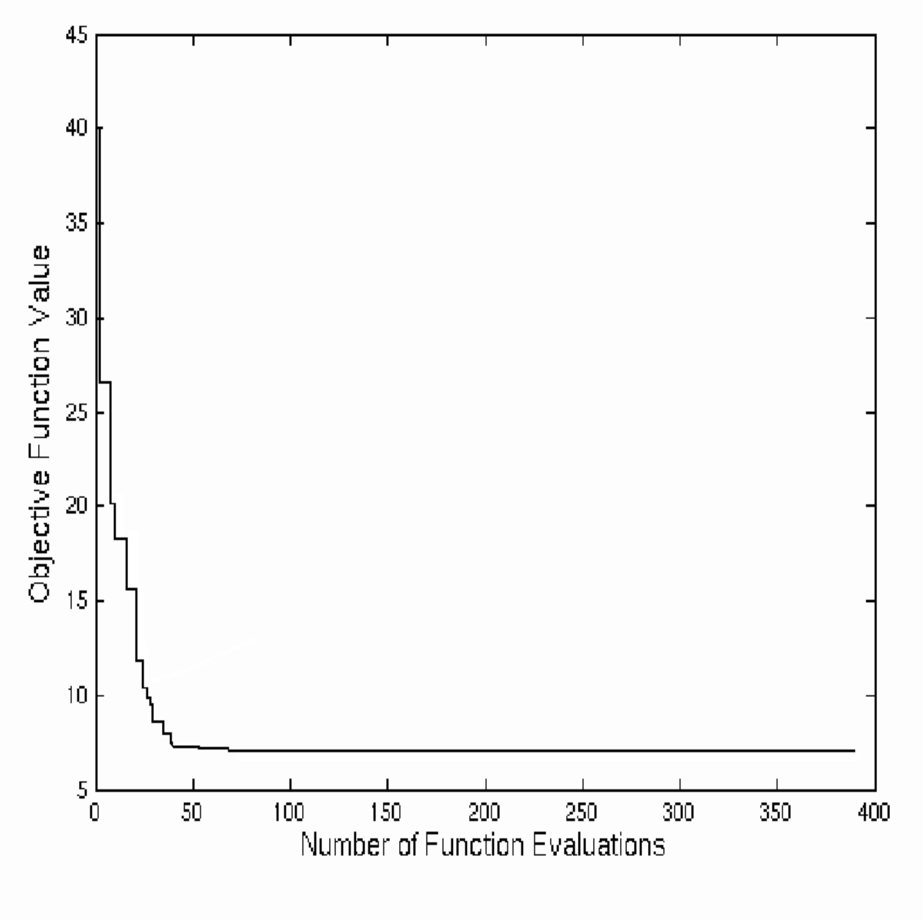

Figure 6.1 Graphical depiction of the sequence of MADS iterations for the three-state example.

6.3.2 Five-State Environment

We now present an example using a five-state random environment with infinitesimal generator matrix

$$Q = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 1 \\ 1 & 0 & 0 & 0 & -1 \end{bmatrix}.$$

The box constraints for the environment-dependent arrival and service rates are given by

$$1 \leq \lambda_j \leq 5, \quad j = 1, 2, \ldots, 5 \tag{6.5}$$

$$0 \leq \mu_j \leq 4, \quad j = 1, 2, \ldots, 5, \tag{6.6}$$

and the remaining problem data is summarized in Table 6.3.

Table 6.3 Problem data for 5-state example.

Vector	Environment State				
	1	2	3	4	5
$\boldsymbol{\lambda}_0$	1.0	1.0	1.0	2.0	5.0
$\boldsymbol{\mu}_0$	2.0	2.5	1.0	3.0	4.0
$\boldsymbol{\xi}$	0.5	1.1	1.5	4.0	1.0
$\boldsymbol{\alpha}$	2.0	0.5	8.5	4.5	6.0
$\boldsymbol{\theta}$	1.0	4.0	2.0	8.0	5.0

Table 6.4 summarizes the optimal solution vectors and objective function value for this illustration. The rate of convergence seen in Figure 6.2 is similar to that of the three-state model. Again, the optimal vectors $\boldsymbol{\lambda}^*$ and $\boldsymbol{\mu}^*$ lie almost exclusively on the boundaries of the region defined by the box constraints in (6.5) and (6.6), with the exception of λ_5, which is set at its upper bound, and thus $\lambda_5^* \geq 1$ turns out to be a nonbinding constraint. Again, this is due to the fact that penalties are not assessed for choosing arrival and service rates on the boundary of the feasible region.

The processing time required to complete the optimization run was much more substantial here than for the model with a three-state random environment. This was a consequence of the exponential increase in the computational effort needed to compute the steady-state probabilities at each objective function evaluation.

Table 6.4 Optimal solution: 5-state example.

Vector	Environment State					$E[W_R]^*$
	1	2	3	4	5	
$\boldsymbol{\lambda}^*$	1.0000	1.0000	1.0000	1.0000	4.9998	13.279
$\boldsymbol{\mu}^*$	3.9996	3.9229	3.9999	3.9999	4.0000	

The availability of the steady-state distribution in numerical form allows us to obtain approximately optimal operating parameters for the M/M/1 retrial queueing system. We have demonstrated that the computational method outlined in this chapter can be applied to a variety of complex queueing systems, so long as the limiting probabilities can be obtained. This includes the large class of queueing models to which the matrix-analytic techniques may be applied. While many other aspects of the optimization of unreliable retrial queueing systems should be investigated (e.g., admission control, prioritized retrial orbit, etc.), our main purpose here was to illustrate the feasibility of using the main results to determine operating parameters (namely the arrival and service rates) that minimize the steady-state mean time in orbit. In the final chapter, we review the main contributions and conclusions of this research and provide some important directions for future work.

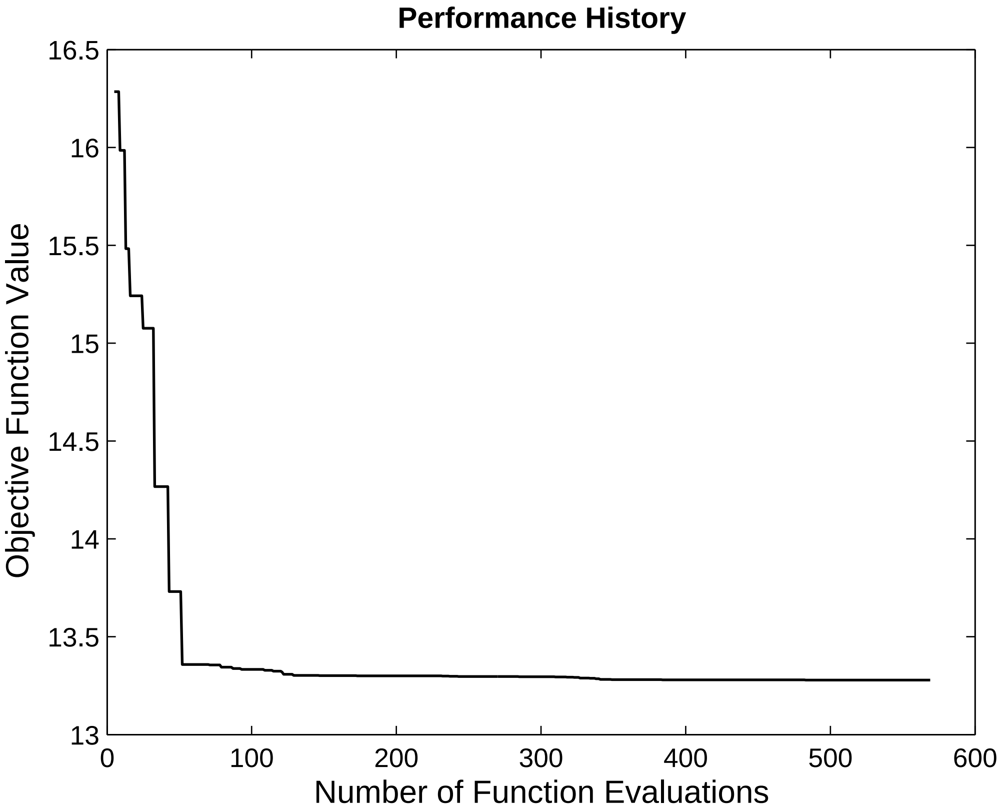

Figure 6.2 Graphical depiction of the sequence of MADS iterations for 5-state example.

7. Contributions and Future Research

In this dissertation, much progress has been made toward the understanding of modulated retrial queueing models subject to breakdowns. Although a number of researchers have studied the steady-state performance of retrial queues, this dissertation is the first to consider Markov modulation *and* server breakdowns in a single model. Moreover, matrix-analytic methods were employed in the mathematical characterization of the stability of the $M/M/1$ and $M/G/1$ versions, as well as to their subsequent numerical solution.

Chapter 4 first considered the foundational case of a single-server system with Poisson arrivals and Markov modulation subject to server failures. For this queueing model, a formal stability analysis was provided that has further implications for more general level-dependent quasi-birth-and-death (LDQBD) processes. By employing classical techniques and the matrix-analytic framework, a computable traffic intensity was derived as an explicit function of the Markov-modulated arrival, service, retrial, failure and repair rates. We further illustrated (in Chapter 6) how the results of this chapter can be used to pragmatically select arrival and service rates that minimize the steady-state mean time a customer spends in orbit, an important measure in a number of applications including communications and computer networks as well as customer contact centers.

We then set out in Chapter 5 to devise similar conditions for the stability of the same retrial queueing system when each customer brings a generally distributed service requirement. In this model, the service distribution is not modulated by the random environment, but all other processes operate exactly as before. In this case, it was shown that the Markov chain embedded at epochs in which the server enters the up and idle state is a Markov chain of $M/G/1$ type. Consequently, we were able to prove the conditions needed for stability of the queueing system. Furthermore, through some tedious mathematics, we were able to explicitly define the transition probability matrix of the embedded Markov chain and define all of its elements in the matrix-analytic format. To the author's knowledge, such results have appeared neither in the stochastic modeling nor the retrial queueing literature.

While this dissertation contributes several basic results for unreliable retrial queueing systems that operate in a random environment, there are several potentially fruitful avenues for extension of this research. One obvious extension is to consider the exponential model with multiple servers. Such a queueing system can be used to model the operation of customer contact centers which have recently received a great deal of attention in the stochastic modeling community. In such case, one may consider customers who also enter the retrial orbit by abandoning the system if their wait time exceeds some predetermined threshold. It may also be instructive to consider non-exponential inter-retrial times since customers may retry the system at arbitrary random intervals.

Another important extension of this work is the consideration of networks of unreliable retrial queueing systems in a random environment. These may, for example, be used to model Internet traffic wherein the various environment states may correspond to the traffic in the network which impacts the performance of each individual queueing station. Analyzing a stochastic network of this kind promises to be very challenging, though worthwhile.

The optimization portion of this research served the purpose of illustrating how the results might be used to improve the performance of a retrial queueing system subject to time-varying conditions. It will be instructive in the future to consider budget constrained systems which add significant complexity to the optimization model of Chapter 6. Furthermore, dealing with a "black box" objective function presents numerical challenges that need to be addressed since it is difficult to discern when a global optimal solution to the problem exists. The computational complexity involved in evaluating candidate solutions is not a minor issue. New algorithms for efficiently computing the approximate steady-state orbit size distribution will be needed to help address this issue.

Appendix A.

A.1 Matrix Binary Operators

We review here some basic, but important, operations on the space of matrices $\mathbb{R}^{m\times n}$ that we shall require in order be able to apply the matrix-analytic methods to the study of quasi-birth-and-death processes (QBDs), the $M/G/1$-type Markov chains, and Markov modulation. We shall first discuss Kronecker products and sums, which are crucial to the manipulation of matrix expressions under integration. We also discuss another binary operator termed the matrix convolution product, which is the matrix version of the convolution of scalar distribution functions.

A.1.1 The Kronecker Product and Sum of Matrices

A useful tool in the representation of block-matrix expressions is the binary operator $\otimes : \mathbb{R}^{m\times n} \times \mathbb{R}^{p\times q} \to \mathbb{R}^{(mp)\times(nq)}$, which is known as the *Kronecker product*, where $\mathbb{R}^{m\times n}$ denotes the set of $m \times n$ matrices over the field of real numbers. The Kronecker product is also known as the *tensor product* when used in a more general algebraic or topological context. The operation is defined as follows:

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B. \end{bmatrix} \tag{1.1}$$

Its usefulness in the analysis of stochastic models derives from the ease with which one may describe the state space of certain juxtaposed stochastic processes (c.f. Theorem 5.3.5 of [82]). Finally, we introduce the *Kronecker sum* $\oplus$ for square matrices, which is defined as

$$A \oplus B \equiv A \otimes I_B + B \otimes I_A,$$

where I_A and I_B are the identity matrices of the same dimensions as that of A and B, respectively.

Some important properties of the Kronecker product bear mentioning since they are crucial to the integration of expressions containing matrices. We begin with the identity

$$(A \otimes C)(B \otimes D) \;=\; AB \otimes CD, \tag{1.2}$$

where all terms are rectangular matrices for whom the products AB and CD are defined. As in [82: p 245], we observe that

$$\exp(A) \otimes \exp(B) \;=\; \exp(A \otimes I + I \otimes B) \;=\; \exp(A \oplus B), \tag{1.3}$$

which is analogous to the behavior of products of scalar exponential terms.

A.1.2 Matrix Convolution Products

One other matrix operation bears mentioning since it will appear in our discussion of $M/G/1$-type processes in Chapter 4. The *matrix convolution product* is, as its name suggests, related to the notion of the convolution of two scalar mass-functions. First, the matrix convolution $(\cdot, \cdot, *)$ of two matrices of distribution or mass-functions, which we shall denote by $F(x) = [f_{ij}(x)]$ and $G(x) = [g_{ij}(x)]$, is defined as the $m \times m$ matrix

$$[F(x) * G(x)]_{ij} \;=\; \sum_{k=1}^{m} \int_0^t f_{ik}(x - x_1) g_{kj}(x_1)\, dx_1 \;=\; \sum_{k=1}^{m} f_{ik}(x) * g_{kj}(x). \tag{1.4}$$

From this definition, we obtain the *Kronecker convolution product* $(\cdot, \cdot, \circledast)$, which is explicitly given by the $m^2 \times m^2$ matrix

$$F(x) \circledast G(x) \;=\; [f_{ij}(x) * G(x)]_{ij}. \tag{1.5}$$

An overview of the operation, related properties, and more advanced mathematical applications may be found in [53].

A.2 PH-Random Variables

Phase-type, or *PH*-distributed random variables define the time-to-absorption of a Markov process with initial probability vector $(\phi_0, \boldsymbol{\phi})$ and with finitely-many transient states, or *phases*, $1, \dots, m$, and absorbing state $m+1$. Processes that behave accordingly are absorbed with probability one. Important applications reside in such fields as reliability theory, where the absorbing state frequently represents equipment failure, after which the process ends or it may be restarted by a repair or replacement. The restarting of the process may be considered to be regenerative (or at least Markov regenerative in the case that the process restarts with different parameters). This trait enables PH-distributions to aptly describe the Markov-modulated Poisson process, which features highly in the queueing model of this dissertation.

The idea of a PH-distribution is an extension of the earlier *method of stages*, which was first studied by Erlang using the distribution that bears his name. In fact, the family of Erlang and hyperexponential distributions are themselves special manifestations of PH-distributions, which in turn are dense in the space of all nondefective distributions. This fact is useful to the approximation of arbitrary (nondefective) distributions. Moreover, the PH-distribution function possesses attractive characteristics that often culminate in computationally tractable mathematical expressions. For these reasons, PH-distributions and their matrix-analytic generalizations are becoming increasingly important in the field of stochastic operations research.

A.2.1 Continuous PH-Distributions

To a continuous PH-random variable X we associate the infinitesimal generator

$$Q = \begin{bmatrix} 0 & \mathbf{0} \\ \boldsymbol{t} & T \end{bmatrix}, \tag{1.6}$$

of a Markovian process, where $\boldsymbol{t}$ is a column vector and T is a square matrix such that $T\boldsymbol{e}+\boldsymbol{t} = \mathbf{0}$, both of dimension m, to the process with non-absorbing states $\{1, \ldots, m\}$ and absorbing state 0. We shall assume for the remainder of this discussion that the ϕ_0 component of the initial probability vector $(\phi_0, \boldsymbol{\phi})$ is equal to zero, since this is equivalent to stating that the corresponding PH-process does not begin in state 0.

Define the matrix exponential $\exp(A)$ to be a function whose domain consists of square matrices A such that

$$\exp(A) \equiv \sum_{k=0}^{\infty} \frac{A^k}{k!}.$$

We may now define the c.d.f. of the PH-distribution associated to the generator Q as

$$F(x) \;=\; P\{X \leq x\} \;=\; 1 - \boldsymbol{\phi} \exp{(Tx)}\boldsymbol{e},$$

which we may, using the fact that $\boldsymbol{t} = -T\boldsymbol{e}$, differentiate in order to obtain the density function of X

$$f(x) \;=\; \boldsymbol{\phi} \exp(Tx)\boldsymbol{t}.$$

It is clear from these definitions that the PH-random variable is completely determined by the stochastic vector $\boldsymbol{\phi}$ and the matrix T. Hence, we call $[\boldsymbol{\phi}, T]$ the *representation* of the PH-distribution that we now denote by $PH(\boldsymbol{\phi}, T)$. As noted by Neuts [83], Latouche and Ramaswami [66], and others, this representation is by no means unique, but one can always find (for a nondefective PH-distribution) a representation such that the probability of absorption from *any* phase $i = 1, \ldots, m$

is equal to one. Such a representation is necessary and sufficient for the matrix T to be invertible (see [66]), and hence, we may always find a representation $[\boldsymbol{\phi}, T]$ for a nondefective PH-distribution for which T is invertible.

A.2.2 Discrete PH-Distributions

Discrete PH-distributions represent the distribution of the time-to-absorption of a discrete-time Markov chain (DTMC). They may also be represented by the pair $[\boldsymbol{\tau}, T]$, except that the matrix portion T of the representation contains the *probability* of transitions between nonabsorbing states of an irreducible DTMC. In other words, the transition probability matrix of the DTMC is given by

$$P = \begin{bmatrix} 0 & \mathbf{0} \\ \boldsymbol{t} & T \end{bmatrix}, \tag{1.7}$$

where $\boldsymbol{t} + T\boldsymbol{e} = \boldsymbol{e}$. Suppose that $X \sim PH_d(\tau, T)$. Since we have

$$P^k = \begin{bmatrix} 0 & \mathbf{0} \\ \boldsymbol{e} - T^k\boldsymbol{e} & T^k \end{bmatrix},$$

the mass function and distribution function of X are given by

$$\begin{aligned} P\{X = 0\} &= \tau_0 & & \\ P\{X = k\} &= \boldsymbol{\tau} T^{k-1}\boldsymbol{t}, & & k \geq 1 \\ P\{X \leq k\} &= 1 - \boldsymbol{\tau} T^k \boldsymbol{e}, & & k \geq 0, \end{aligned}$$

(see [66: Thm 2.5.3]). Most importantly, nondefective discrete PH-distributions are also guaranteed the existence of a representation such the $I - T$ is nonsingular; in other words, the probability of absorption from *any* state is equal to 1.

A.2.3 Properties of the PH-Distribution

Since there will be occasion to use PH-distributions in conditioning arguments, we will give some additional properties.

Lemma A.1. *Let $[\boldsymbol{\phi}, T]$ be the representation of a nondefective continuous PH-distribution. Then the following holds:*

$$\int_0^\infty \exp(Tx)\,dx \;=\; (-T)^{-1}. \tag{1.8}$$

Proof. By definition,

$$\begin{aligned} F(x) \;&=\; \int_0^x f(u)\,du \\ &=\; \boldsymbol{\phi}\int_0^x \exp(Tu)\,du\,(-T\boldsymbol{e}). \end{aligned} \tag{1.9}$$

Taking the limit of (1.9) as $x \to \infty$ gives us the following relationship:

$$1 \;=\; \boldsymbol{\phi}\int_0^\infty \exp(Tu)\,du\,(-T\boldsymbol{e}). \tag{1.10}$$

Since the distribution is assumed to be nondefective, we may assume that T is nonsingular. We thus obtain our result by substituting $\boldsymbol{\phi I e}$ for 1 in (1.10) and solving. □

The convolution of two independent PH-random variables X and Y will also be required. The following appears as Theorem 2.6.1 in [66]:

Theorem A.1. *Let $X \sim PH(\boldsymbol{\tau}, T)$ and $Y \sim PH(\boldsymbol{\beta}, B)$ be independent with m and n phases, respectively. Their sum $X + Y \sim PH(\boldsymbol{\gamma}, C)$ with $m + n$ phases, where*

$$\boldsymbol{\gamma} = [\boldsymbol{\tau}, \tau_0\boldsymbol{\beta}] \qquad \textit{and} \qquad C = \begin{bmatrix} T & \boldsymbol{t\beta} \\ 0 & S \end{bmatrix}.$$

A.3 Markov-Modulated Poisson Processes (MMPP)

The MMPPs are members of a class of processes termed *doubly-stochastic*, the name taking its cue from the fact that the *parameters* of the statistical distribution of the random variables of the process are themselves random variables. MMPPs are considered a special case due to the fact that the exponential distribution of its interarrivals has a parameter that itself randomly varies according to a finite-state Markov chain. In other words, if the modulating Markov process $\{Z(t) : t \geq 0\}$ has the state space $S = \{1, 2, \ldots, m\}$, whenever $Z(t) = j$, arrivals occur acording to a Poisson process with rate λ_j, $j \in S$. An excellent and highly-useful overview of the MMPP and its properties may be found in [40].

A.3.1 The Markov-Modulated Poisson Process as a Markov Renewal Process

It is a well-established fact that the standard Poisson process $\{N(t) : t \geq 0\}$ is a *renewal* process characterized by the i.i.d. nature of the interarrival times. However, as in [40, 82], an MMPP with exponential rates contained in the vector $\boldsymbol{\eta}$ is *not* a standard renewal process due to the dependence of the interarrival durations upon the evolution of the modulating process. Nevertheless, it is still possible to describe the MMPP instead as a *Markov* renewal process or *semi-Markov process* for which we require only that the increments of the process behave in accordance with the Markov property. In order to see this, consider the MMPP at arrival epochs $t_n \geq 0$ and let the state of the modulating process at t_n be $Z_n = Z(t_n^+)$ and $\tau_n = t_n - t_{n-1}$. Then the process $\{(Z_n, \tau_n) : n \geq 0\}$ defines a Markov renewal sequence. In other words, since we embed the sequence at transitions of a Markov process, the Markov property may be said to hold for the increments of the embedded sequence. Most importantly, it is a well-known property of Markov renewal sequences that $\{Z_n : n \geq 0\}$ itself defines a discrete-time Markov chain.

We next introduce the *kernel* $K(x)$ of the semi-Markov process, which is defined to be the matrix with entries given by

$$K_{ij}(x) \;=\; P\left\{Z_n = j, \tau_n \leq x \,|\, Z_{n-1} = i\right\}, \quad x \geq 0, \;\; n \geq 1,$$

where i and j belong to the state space of the Markov chain $\{Z_n : n \geq 0\}$. The importance of the kernel lies in the fact that $P = K(\infty)$ is the transition probability matrix of the MMPP embedded at arrival times. Thus, in utilizing Markov renewal theory, we will have enabled the use of elementary methods defined for Markov chains in the analysis of the steady-state conditions of a non-Markovian process. We will use this property later in constructing an embedded Markov process that will serve as a proxy for the analysis of a non-Markovian $M/G/1$-type queueing system.

A.3.2 The Relationship of the PH-Distributions to the Markov-Modulated Poisson Process

The utility of the PH-distribution for our queueing model rests in its association to the MMPP. As mentioned previously, inter-event times in a MMPP may no longer be characterized by a single exponential distribution. Neuts [83] has shown that Markovian processes whose rate parameters $\boldsymbol{\eta}$ vary lexicographically according to the states of an external Markov chain with generator Q possess inter-event times distributed as $PH(\boldsymbol{\phi}), T_\eta$, where

$$T_\eta \;=\; Q - \Delta(\boldsymbol{\eta})\,.$$

Thus, the distribution of time between the arrival of primary customers to the retrial queue is $PH(\boldsymbol{\phi}_\lambda, T_\lambda)$ for a suitable choice of initial probability vector $\boldsymbol{\phi}_\lambda$. In this light, it is clear that the MMPP is a *PH-renewal process*, which belongs to a class of point processes termed *versatile* by Neuts and, more recently, the *Markovian arrival process* (MAP) or the *batch Markovian arrival process* (BMAP).

Armed with this knowledge, it is now possible to fully describe the counting process associated to the MMPP. Define $N(t)$ to be the number of renewals (i.e. arrivals) in $(0, t]$, and set

$$\Phi^{\boldsymbol{\eta}}_{jj'}(\nu, t) \; = \; P\{N(t) = \nu, Z(t) = j' \,|\, N(0) = 0, Z(0) = j\}, \tag{1.11}$$

where j and j' are in the state space of the modulating Markov chain $\{Z(t) : t \geq 0\}$ and $\nu \in \mathbb{Z}^+$. As in [40], (1.11) obeys the forward Chapman-Kolmogorov equations

$$\Phi^{\boldsymbol{\eta}'}(0, t) \;\; = \;\; \Phi^{\boldsymbol{\eta}}(0, t)T_\eta, \tag{1.12}$$

$$\Phi^{\boldsymbol{\eta}'}(k, t) \;\; = \;\; \Phi^{\boldsymbol{\eta}}(k, t)T_\eta - \Phi^{\boldsymbol{\eta}}(k-1, t)(T_\eta \boldsymbol{e} \boldsymbol{\phi}_\eta), \quad k \geq 1, \tag{1.13}$$

which, of course, directly implies that

$$\Phi^{\boldsymbol{\eta}}(0, t) \; = \; \exp(T_\eta x).$$

Equations (1.12) and (1.13) may also be used to derive the probability generating function $\Phi^{\boldsymbol{\eta}*}$, which is stated in terms of the following lemma:

Lemma A.2. *Consider the MMPP with the exponential arrival rate parameters contained in the vector* $\boldsymbol{\eta}$ *and infinitesimal generator* Q *of the random environment. If* $t > 0$ *is a real number, then the probability generating function* $\Phi^{\boldsymbol{\eta}*}(z, t)$ *of* $\Phi^{\boldsymbol{\eta}}(\nu, t)$ *is given by*

$$\Phi^{\boldsymbol{\eta}*}(z, t) \; = \; \exp[Q - (1 - z)\Delta(\boldsymbol{\eta})], \tag{1.14}$$

where z *is a number that resides in the interval* $(0, 1)$*.*

We shall require Lemma A.2 in constructing the entries of the semi-Markov kernel $Q^*(x)$ for the queueing system that we shall analyze in Section 5.2.

For completeness, we consider the value of the initial probability vector of a generic MMPP with rate vector $\boldsymbol{\eta}$. The choice of this vector determines the start time of the MMPP, as well as the state in which the process begins. There is no restriction on the choice of the initial probability vector, save that its elements must sum to unity. However, there are two commonly used methods in determining its value. The first is the selection that results in the *environment-stationary* version of the MMPP. This corresponds to the selection of the steady-state probability vector $\boldsymbol{p}$ of the random environment, which subsequently starts the MMPP at some point in time at which the environment achieves equilibrium. The other version of the MMPP is known as the *time-stationary* version. Here, the initial probability vector is chosen in such a way that the Markov renewal sequence corresponding to the MMPP is itself in equilibrium. As it turns out, both versions are stochastically equivalent (see [82: Thm 5.3.3]).

In characterizing the semi-Markov kernel $Q^*(x)$ for queueing systems with more than one MMPP (or PH) arrival stream, it becomes necessary to condition upon the order of arrivals in a regenerative cycle. For this purpose, it is necessary to determine $\prod_{j=1}^{N} P\{X_i \leq X_j,\ i \in \{1, \ldots, N\}$, for a finite set of PH-random variables with representations $[\boldsymbol{\phi}_i, T_i]$. Suppose that $N = 2$ and that $F_i(x) = 1 - \boldsymbol{\phi}_i \exp(Tx)\boldsymbol{e}$ are the c.d.f.s corresponding to the random variables $X_1, \ldots, X_N$. We compute these probabilities by conditioning upon the values of $F_2(x)$, noting that the dimension of $\boldsymbol{e}$, the column vector of ones, varies accordingly, even though this is not explicitly indicated.

$$\begin{aligned}
P\{X_1 \leq X_2\} &= \int_0^\infty F_1(x)\, dF_2(x) \\
&= \int_0^\infty (1 - \boldsymbol{\phi}_1 \exp(T_1 x)\boldsymbol{e})\,(-\boldsymbol{\phi}_2 \exp(T_2 x) T_2 \boldsymbol{e})\ dx \\
&= 1 + (\boldsymbol{\phi}_1 \otimes \boldsymbol{\phi}_2)\left[\int_0^\infty \exp(T_1 x) \otimes \exp(T_2 x)\, dx\right](\boldsymbol{e} \otimes T_2 \boldsymbol{e}) \\
&= 1 + (\boldsymbol{\phi}_1 \otimes \boldsymbol{\phi}_2)\left[\int_0^\infty \exp[(T_1 \oplus T_2)x]\, dx\right](\boldsymbol{e} \otimes T_2 \boldsymbol{e})
\end{aligned}$$

$$= 1 + (\boldsymbol{\phi}_1 \otimes \boldsymbol{\phi}_2)\,[-(T_1 \oplus T_2)]^{-1}\,(\boldsymbol{e} \otimes T_2\boldsymbol{e}). \tag{1.15}$$

The equality

$$\begin{aligned} &\int_0^\infty \boldsymbol{\phi}_1 \exp(T_1 x)\boldsymbol{\phi}_2 \exp(T_2 x) T_2 \boldsymbol{e}\, dx \\ &\qquad = \;(\boldsymbol{\phi}_1 \otimes \boldsymbol{\phi}_2)\left[\int_0^\infty \exp(T_1 x) \otimes \exp(T_2 x)\, dx\right](\boldsymbol{e} \otimes T_2\boldsymbol{e}) \end{aligned}$$

follows from first noticing that the ordinary product of two scalars is also a tensor product, to which we may then apply the product rule (1.2). In other words,

$$\begin{aligned} &\boldsymbol{\phi}_1 \exp(T_1 x)\boldsymbol{e}\boldsymbol{\phi}_2 \exp(T_2 x) T_2 \boldsymbol{e} \\ &\qquad = (\boldsymbol{\phi}_1 \exp(T_1 x)\boldsymbol{e}) \otimes (\boldsymbol{\phi}_2 \exp(T_2 x) T_2 \boldsymbol{e}) \\ &\qquad = (\boldsymbol{\phi}_1 \otimes \boldsymbol{\phi}_2)(\exp(T_1 x) \otimes \exp(T_2 x))(\boldsymbol{e} \otimes T_2\boldsymbol{e}) \\ &\qquad = (\boldsymbol{\phi}_1 \otimes \boldsymbol{\phi}_2) \exp[(T_1 \oplus T_2)x](\boldsymbol{e} \otimes T_2\boldsymbol{e}), \end{aligned} \tag{1.16}$$

with the final equality resulting from the identity (1.3).

In considering the MMPP in the context of the $MMPP/G/1$ queue, we obtain a Markov renewal process by considering the number of arrivals to the queue at instants $S_n = T_n^+$ just after the nth departure. Let Y_n be the size of the system at T_n and let $\tau_n = S_n - S_{n-1}$. In this way, we obtain the Markov renewal sequence $\{(Y_n, Z_n, \tau_n) : n \geq 0\}$, which defines the kernel matrix $Q^*(x)$ consisting of the block elements

$$A_{jj'}(x) \;=\; [Q^*_{ii'}(x)]_{jj'} = P\,\{Y_n = i',\, Z_n = j',\, \tau_n \leq x \,|\, Y_{n-1} = i,\, Z_{n-1} = j\}$$

if $i > 0$ and

$$B_{jj'}(x) \;=\; [Q^*_{0i'}(x)]_{jj'} = P\left\{Y_n = i',\, Z_n = j',\, \tau_n \leq x \,|\, Y_{n-1} = 0,\, Z_{n-1} = j\right\},$$

where j, j', and x are defined as before and i and i' are nonnegative integers that correspond to the number in the system. The matrix $P = B(\infty)$ is the transition probability matrix of the Markov chain embedded just after departure instants. The kernel $Q^*(x)$ for $M/G/1$-type processes assumes a distinctive structure which, like the tridiagonal form of the QBD generator, a set of class-based methods that may be applied to a wide spectrum of related models.

Bibliography

1. Abramson, M. A. NOMADm Optimization Software. Website. URL `http://www.afit.edu/en/ENC/Faculty/MAbramson/NOMADm.html`.

2. Abramson, M. A. and Audet, C. (2006). Convergence of mesh adaptive direct search to second-order stationary points. *SIAM Journal on Optimization*, **17**(2), 606–619.

3. Adan, I. J. B. F. and Kulkarni, V. G. (2003). Single-server queue with Markov-dependent inter-arrival and service times. *Queueing Systems: Theory and Applications*, **45**(2), 113–134.

4. Aissani, A. (1988). On the $M/G/1/1$ queueing system with repeated orders and unreliable server. *Journal of Technology*, **6**, 98–123. (in French).

5. Aissani, A. (1993). Unreliable queuing with repeated orders. *Microelectronics and Reliability*, **33**(14), 2093–2106.

6. Aissani, A. (1994). A retrial queue with redundancy and unreliable server. *Queueing Systems: Theory and Applications*, **17**(3-4), 431–449.

7. Aissani, A. and Artalejo, J. R. (1998). On the single server retrial queue subject to breakdowns. *Queueing Systems: Theory and Applications*, **30**(3-4), 309–321.

8. Akar, N. and Sohraby, K. (1997). An invariant subspace approach in $M/G/1$ and $G/M/1$ type Markov chains. *Communications in Statistics.Stochastic Models*, **13**(3), 381–416.

9. Almasi, B., Roszik, J., and Sztrik, J. (2005). Homogeneous finite-source retrial queues with server subject to breakdowns and repairs. *Mathematical & Computer Modelling*, **42**(5/6), 673–682.

10. Anisimov, V. and Sztrik, J. (1989). Asymptotic analysis of some complex renewable systems operating in random environments. *European Journal of Operational Research*, **41**(2), 162–168.

11. Anisimov, V. V. and Atadzhanov, K. L. (1994). Diffusion approximation of systems with repeated calls and an unreliable server. *Journal of Mathematical Sciences*, **72**(2), 3032–3034.

12. Asmussen, S. (1991). Ladder heights and the Markov-modulated $M/G/1$ queue. *Stochastic Processes and their Applications*, **37**(2), 313–326.

13. Asmussen, S. (2000). Matrix-analytic models and their analysis. *Scandinavian Journal of Statistics*, **27**, 193.

14. Atencia, I. and Moreno, P. (2006). A discrete-time $Geo/G/1$ retrial queue with the server subject to starting failures. *Annals of Operations Research*, **141**(1-4), 85–107.

15. Audet, C. and J.E. Dennis, J. (2006). Mesh adaptive direct search algorithms for constrained optimization. *SIAM Journal on Optimization*, **17**(1), 188–217.

16. Audet, C. and Orban, D. (2006). Finding optimal algorithmic parameters using derivative-free optimization. *SIAM Journal on Optimization*, **17**(3), 642–664.

17. Avi-Itzhak, B. (1963). Preemptive repeat priority queues as a special case of the multipurpose server problem–I. *Operations Research*, **11**(4), 597–609.

18. Avi-Itzhak, B. (1963). Preemptive repeat priority queues as a special case of the multipurpose server problem–II. *Operations Research*, **11**(4), 610–619.

19. Avi-Itzhak, B. and Naor, P. (1963). Some queuing problems with the service subject to breakdown. *Operations Research*, **11**(3), 303–320.

20. Baccelli, F. and Makowski, A. A. (1986). Stability and bounds for single server queues in random environment. Technical report, Inst. Nat. Recherche Inf. Autom., Le Chesnay, France.

21. Bhat, V. N. (1995). A queueing model in an alternating random environment. *Computers & Industrial Engineering*, **28**(2), 323–328.

22. Bini, D. A., Meini, B., and Ramaswami, V. (1998). Analyzing $M/G/1$ paradigms through QBDs: the role of the block structure in computing the matrix G. In A. S. Alfa and S. R. Chakravarthy, editors, *Advances in Matrix Analytic Methods for Stochastic Models*. Notable Publications, Inc., Neshanic Station, NJ, 73–86.

23. Borst, S., Boucherie, R. J., Boxma, O. J., Key, P., and Smith, D. (1999). ERMR: A generalised equivalent random method for overflow systems with repacking. In P. Key and D. Smith, editors, *Teletraffic Engineering in a Competitive World. Proceedings of the International Teletraffic Congress - ITC-16.*, volume 3a. Elsevier Science; Alcatel, Amsterdam, the Netherlands, 313–323.

24. Bourgin, R. D. and Cogburn, R. (1981). On determining absorption probabilities for Markov chains in random environments. *Advances in Applied Probability*, **13**(2), 369–387.

25. Bright, L. and Taylor, P. G. (1995). Calculating the equilibrium distribution in level dependent quasi-birth-and-death processes. *Communications in Statistics: Stochastic Models*, **11**(3), 497–525.

26. Burman, D. Y. and Smith, D. R. (1986). An asymptotic analysis of a queueing system with Markov-modulated arrivals. *Operations Research*, **34**(1), 105–119.

27. Chang, C.-S. and Pinedo, M. (1990). Bounds and inequalities for single-server loss systems. *Queueing Systems*, **6**, 425–436.

28. Chen, F. and Song, J.-S. (2001). Optimal policies for multiechelon inventory problems with Markov-modulated demand. *Operations Research*, **49**(2), 226–234.

29. Choi, B. D. and Park, K. K. (1990). The $M/G/1$ retrial queue with Bernoulli schedule. *Queueing Systems: Theory and Applications*, **7**, 219–228.

30. Clos, C. (1948). An aspect of the dialing behaviour of subscribers and its effect on the trunk plant. *Bell Systems Technical Journal*, **27**, 424–445.

31. Cohen, J. W. (1957). An aspect of the dialing behaviour of subscribers and its effect on the trunk plant. *Philips Telecommunication Review*, **18**(2), 49–101.

32. Cohen, J. W. (1957). The full availability group of trunks with an arbitrary distribution of the inter-arrival times and a negative exponential holding time distribution. *Simon Stevin: A Quarterly Journal of Pure and Applied Mathematics*, **31**, 169–181.

33. Dohi, T., Kaio, N., and Osaki, S. (2001). Optimal periodic maintenance strategy under an intermittently used environment. *IIE Transactions*, **33**(12), 1037.

34. Dudin, A. N. and Klimenok, V. I. (1997). Calculation of the characteristics of a single-server system functioning in a synchronous Markovian random environment. *Rossiĭskaya Akademiya Nauk.Avtomatika i Telemekhanika*, (1), 74–84.

35. Ebrahimi, N. (2006). System reliability based on system wear. *Stochastic Models*, **22**(1), 21–36.

36. Economou, A. (2004). Stationary distributions of discrete-time Markov chains in random environment: exact computations and bounds. *Stochastic Models*, **20**(1), 103–127.

37. Eisen, M. and Tainiter, M. (1963). Stochastic variations in queuing processes. *Operations Research*, **11**(6), 922–927.

38. Falin, G. (1990). A survey of retrial queues. *Queueing Systems: Theory and Applications*, **7**(2), 127–167.

39. Falin, G. I. and Artalejo, J. R. (1998). A finite source retrial queue. *European Journal of Operational Research*, **108**(2), 409.

40. Fischer, W. and Meier-Hellstern, K. (1992). Markov-modulated Poisson process (MMPP) cookbook. *Performance Evaluation*, **18**(2), 149–171.

41. Georgiadis, L., Neely, M. J., and Tassiulas, L. (2006). Resource allocation and cross-layer control in wireless networks. *Foundations and Trends in Networking*, **1**(1), 1–144.

42. Gharbi, N. and Ioualalen, M. (2006). GSPN analysis of retrial systems with server breakdowns and repairs. *Applied Mathematics and Computation*, **174**(2), 1151–1168.

43. Golovko, N. I. and Korotaev, I. A. (1990). Queuing systems with a randomly varying rate of the input flow. *Automation and Remote Control*, **51**(7), 921–926.

44. Harris, C. M. (1967). Queues with state dependent stochastic service rates. *Operations Research*, **15**(1), 117–130.

45. Hinderer, K. and Waldmann, K.-H. (2001). Cash management in a randomly varying environment. *European Journal of Operational Research*, **130**(3), 468–485.

46. Hu, D. (2005). The decomposition of state space for Markov chain in random environment. *Acta Mathematica Scientia. Series B.English Edition*, **25**(3), 555–568.

47. Kalpakam, S. and Arivarignan, G. (1989). A lost sales inventory system in a random environment. *Stochastic Analysis and Applications*, **7**(4), 367–385.

48. Keilson, J., Cozzolino, H., and Young, A. (1968). A service system with unfilled requests repeated. *Operations Research*, **16**, 1126–1137.

49. Keilson, J. and Kooharian, A. (1960). On time dependent queuing processes. *Annals of Mathematical Statistics*, **31**, 104–112.

50. Kella, O. and Whitt, W. (1992). A storage model with a two-stage random environment. *Operations Research*, **40**(3), 257. Supplement 2.

51. Kharoufeh, J. P. (2003). Explicit results for wear processes in a Markovian environment. *Operations Research Letters*, **31**(3), 237–244.

52. Kharoufeh, J. P. and Cox, S. M. (2005). Stochastic models for degradation-based reliability. *IIE Transactions*, **37**(6), 533–542.

53. Kilicman, A. and Zhour, Z. A. A. A. (2007). Kronecker operational matrices for fractional calculus and some applications. *Applied Mathematics and Computation*, **187**(1), 250–265.

54. Klimenok, V. (2005). A $BMAP/SM/1$ queueing system with hybrid operation mechanism. *Automation & Remote Control*, **66**(5), 779–790.

55. Kodera, T. and Miyazawa, M. (2002). An $M/G/1$ queue with Markov-dependent exceptional service times. *Operations Research Letters*, **30**(4), 231.

56. Kogan, Y. A. and Litvin, V. G. (1976). Computing the characteristics of a queueing system with a finite buffer and operating in a random environment. *Automation and Remote Control*, **37**(12), 1828–1835.

57. Korotaev, I. A. and Spivak, L. R. (1992). Queueing systems in a semi-Markovian random environment. *Automation and Remote Control*, **53**(7), 1028–1033.

58. Kosten, L. (1947). On the influence of repeated calls in the theory of probabilities of blocking. *De Ingenieur*, **59**, 1–25. (in Dutch).

59. Kroese, D. P. and Nicola, V. F. (1999). Efficient estimation of overflow probabilities in queues with breakdowns. *Performance Evaluation*, **36-37**, 471–484.

60. Kulkarni, V. G. (1983). On queueing systems with retrials. *Journal of Applied Probability*, **20**(2), 380–389.

61. Kulkarni, V. G. (1995). *Modeling and Analysis of Stochastic Systems.* Texts in Statistical Science. Chapman & Hall/CRC, 1st edition.

62. Kulkarni, V. G. and Choi, B. D. (1990). Retrial queues with server subject to breakdowns and repairs. *Queueing Systems: Theory and Applications*, **7**(2), 191–208.

63. Kulkarni, V. G. and Liang, H. M. (1997). *Retrial queues revisited.* Frontiers in queueing; Probab. Stochastics Ser. CRC, 19–34.

64. Kumar, K. B. and Arivudainambi, D. (2002). The $M/G/1$ retrial queue with Bernoulli schedules and general retrial times. *Computers & Mathematics with Applications*, **43**(1-2), 15–30.

65. Kumar, K. B., Madheswari, P. S., and Vijayakumar, A. (2002). The $M/G/1$ retrial queue with feedback and starting failures. *Applied Mathematical Modelling*, **26**(11), 1057–1075.

66. Latouche, G. and Ramaswami, V. (1999). *Introduction to Matrix–Analytic Methods in Stochastic Modeling.* ASA–SIAM Series on Statistics and Applied Probability. American Stat. Assoc. and the Soc. for Indust. and Applied Mathematics, Alexandria, VA and Philadelphia, PA.

67. Latouche, G. and Taylor, P. G. (2003). Drift conditions for matrix-analytic models. *Mathematics of Operations Research*, **28**(2), 346.

68. Lee, D.-S. and Li, S.-Q. (1992). Transient analysis of multi-server queues with Markov-modulated Poisson arrivals and overload control. *Performance Evaluation*, **16**(1-3), 49–66.

69. Leese, E. L. and Boyd, D. W. (1966). Numerical methods of determining the transient behaviour of queues with variable arrival rates. *CORS Journal*, **4**(1), 1–13.

70. Li, H. and Yang, T. (1995). A single-server retrial queue with server vacations and a finite number of input sources. *European Journal of Operational Research*, **85**(1), 149.

71. Li, Q.-L., Ying, Y., and Zhao, Y. Q. (2006). A $BMAP/G/1$ retrial queue with a server subject to breakdowns and repairs. *Annals of Operations Research*, **141**(1), 233.

72. Libman, L. and Orda, A. (2002). Optimal retrial and timeout strategies for accessing network resources. *IEEE/ACM Transactions on Networking*, **10**(4), 551–564.

73. Mahabhashyam, S. and Gautam, N. (2005). On queues with Markov-modulated service rates. *Queueing Systems: Theory and Applications*, **51**(1/2), 89–113.

74. Medhi, J. (2003). *Stochastic models in queueing theory.* Academic Press, Amsterdam, second edition.

75. Mitrani, I. (2005). Approximate solutions for heavily loaded Markov-modulated queues. *Performance Evaluation*, **62**(1-4), 117–131.

76. Miyazawa, M. (1985). The intensity conservation law for queues with randomly changed service rate. *Journal of Applied Probability*, **22**(2), 408–418.

77. Mokaddis, G. S., Elias, S. S., and Metwally, S. A. (1984). The characteristics of a bulk service system operating in a random environment. *Ain Shams Science Bulletin: Part A*, **24**, 85–101.

78. Mokaddis, G. S., Elias, S. S., and Metwally, S. A. (1985). The characteristics of a queueing system operating in a random environment. *Bulletin of the Calcutta Mathematical Society*, **77**(2), 115–124.

79. Neuts, M. F. (1968). The joint distribution of the virtual waitingtime and the residual busy period for the $M/G/1$ queue. *Journal of Applied Probability*, **5**, 224–229.

80. Neuts, M. F. (1971). A queue subject to extraneous phase changes. *Advances in Applied Probability*, **3**, 78–119.

81. Neuts, M. F. (1989). The fundamental period of the queue with Markov-modulated arrivals. In T. W. Anderson, K. B. Athreya, and D. L. Iglehart, editors, *Probability, Statistics, and Mathematics: Papers in Honor of Samuel Karlin.* Academic Press, Boston, MA, 187–200.

82. Neuts, M. F. (1989). *Structured stochastic matrices of $M/G/1$ type and their applications*, *Probability: Pure and Applied*, volume 5. Marcel Dekker Inc, New York.

83. Neuts, M. F. (1999). *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach.* Dover Publications, Inc., New York, NY.

84. Neuts, M. F. (1978). The $M/M/1$ queue with randomly varying arrival and service rates. *Opsearch*, **15**(4), 139–157.

85. Neuts, M. F. (1978). Further results on the $M/M/1$ queue with randomly varying rates. *Opsearch*, **15**(4), 158–168.

86. Özekici, S. and Soyer, R. (2004). Reliability modeling and analysis in random environments. In *Mathematical Reliability: An Expository Perspective*, *Internat. Ser. Oper. Res. Management Sci.*, volume 67. Kluwer Academic Publishers, Boston, MA, 249–273.

87. Pakes, A. G. (1969). Some conditions for ergodicity and recurrence of Markov chains. *Operations Research*, **17**, 1058–1061.

88. Purdue, P. (1974). The $M/M/1$ queue in a Markovian environment. *Operations Research*, **22**(3), 562–569.

89. Ramaswami, V. (1998). The generality of quasi birth-and-death processes. In A. S. Alfa and S. R. Chakravarthy, editors, *Advances in Matrix Analytic Methods for Stochastic Models*. Notable Publications, Inc, Neshanic Station, NJ, 93–113.

90. Rao, B. M. and Posner, M. J. M. (1984). On the output of an $M/M/1$ queue with randomly varying system parameters. *Operations Research Letters*, **3**(4), 191–197.

91. 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*, **11**(3), 465–483.

92. Riska, A. and Smirni, E. (2002). Exact aggregate solutions for $M/G/1$-type Markov processes. In *SIGMETRICS '02: Proceedings of the 2002 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*. ACM Press, New York, NY, USA, 86–96.

93. Rolski, T. (1981). Queues with non-stationary input stream: Ross's conjecture. *Advances in Applied Probability*, **13**(3), 603–618.

94. Sengupta, B. (1987). Sojourn time distributions for the $M/M/1$ queue in a Markovian environment. *European Journal of Operational Research*, **32**(1), 140–149.

95. Sennott, L. I., Humblet, P. A., and Tweedie, R. L. (1983). Mean drifts and the non-ergodicity of Markov chains. *Operations Research*, **31**(4), 783–788.

96. Sherman, N. P. and Kharoufeh, J. P. (2006). An $M/M/1$ retrial queue with unreliable server. *Operations Research Letters*, **34**(6), 697–705.

97. Sherman, N. P., Kharoufeh, J. P., and Abramson, M. A. (2006). Analysis and control of an $M/G/1$ retrial queue with unreliable server. Preprint.

98. Shurbet, G., Lewis, T., and Boullion, T. (1974). Quadratic matrix equations. *The Ohio Journal of Science*, **74**(5), 273–277.

99. South, J. B. (1985). Continuous excess capacity versus intermittent extra capacity to control average queue size in a random environment. *Production and Inventory Management*, **26**(1), 103–110.

100. Stanford, D., Horn, W., and Latouche, G. (2006). Tri-layered QBD processes with boundary assistance for service resources. *Stochastic Models*, **22**, 361–382.

101. Sztrik, J. (1993). Asymptotic analysis of a heterogeneous finite-source communication system operating in random environments. *Publicationes Mathematicae Debrecen*, **42**(3-4), 225–238.

102. Sztrik, J., Almasi, B., and Roszik, J. (2006). Heterogeneous finite-source retrial queues with server subject to breakdowns and repairs. *Journal of Mathematical Sciences*, **132**(5), 677–685.

103. Sztrik, J. and Bunday, B. D. (1993). Asymptotic analysis of the heterogeneous machine interference problem with random environments. *Applied Mathematical Modelling*, **17**(2), 89–97.

104. Sztrik, J. and Bunday, B. D. (1993). Machine interference problem with a random environment. *European Journal of Operational Research*, **65**(2), 259–269.

105. Sztrik, J. and Kim, C. S. (2003). Markov-modulated finite-source queueing models in evaluation of computer and communication systems. *Mathematical & Computer Modelling*, **38**(7-9), 961.

106. Sztrik, J. and Lukashuk, L. (1991). Modelling of a communication system evolving in a random environment. *Acta Cybernetica*, **10**(1), 85–91.

107. Tang, Y. H. (1997). A single-server $M/G/1$ queueing system subject to breakdowns–some reliability and queueing problems. *Microelectronics and Reliability*, **37**(2), 315–321.

108. van Leeuwaarden, J. S. H. and Winands, E. M. M. (2006). Quasi-birth-and-death processes with an explicit rate matrix. *Stochastic Models*, **22**.

109. Voevdskiı, E. N. and Postan, M. Y. (1985). Mnogokanalcprime naya sistema massovogo obsluzhivaniya v sluchaui noui srede. (Russian) [A multichannel queueing system in a random environment] [preprint], 85-20.

110. Walrand, J. (1991). *Communication Networks: A First Course*. The Aksen Associates Series in Electrical and Computer Engineering. Richard D. Irwin, Inc., and Aksen Associates, Inc., Homewood, IL and Boston, MA.

111. Wang, J., Cao, J., and Li, Q. (2001). Reliability analysis of the retrial queue with server breakdowns and repairs. *Queueing Systems: Theory and Applications*, **38**(4), 363.

112. Wilkinson, R. I. (1956). Theories for toll traffic engineering in the U.S.A. *Bell Systems Technical Journal*, **35**(2), 421–507.

113. Yang, T. and Templeton, J. G. C. (1987). A survey on retrial queues. *Queueing Systems: Theory and Applications*, **2**(3), 201–233.

114. Yechiali, U. (1973). A queuing-type birth-and-death process defined on a continuous-time Markov chain. *Operations Research*, **21**(2), 604.

115. Yechiali, U. and Naor, P. (1971). Queuing problems with heterogeneous arrivals and service. *Operations Research*, **19**(3), 722–734.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of the collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to an penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)* 13-09-2007	**2. REPORT TYPE** **Doctoral Dissertation**	**3. DATES COVERED** *(From – To)* Sept 2004 – Sept 2007

4. TITLE AND SUBTITLE

UNRELIABLE RETRIAL QUEUES IN A RANDOM ENVIRONMENT

5a. CONTRACT NUMBER
F1ATA06034J001

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Cordeiro, James D., Major, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)
Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hosbson Way, Building 640
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT/DS/ENS/07-03

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Dr. Donald Hearn
Air Force Office of Scientific Research
Suite 325, Room 3112
875 Randolph Street
Arlington, VA 22203-1768 Email: Donald.Hearn@afosr.af.mil; Tel.: 703-696-1142

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This dissertation investigates stability conditions and approximate steady-state performance measures for unreliable, single-server retrial queues operating in a randomly evolving environment. In such systems, arriving customers that find the server busy or failed join a retrial queue from which they attempt to regain access to the server at random intervals. Such models are useful for the performance evaluation of communications and computer networks which are characterized by time-varying arrival, service and failure rates. To model this time-varying behavior, we study systems whose parameters are modulated by a finite Markov process. Two distinct cases are analyzed. The first considers systems with Markov-modulated arrival, service, retrial, failure and repair rates assuming all interevent and service times are exponentially distributed. The joint process of the orbit size, environment state, and server status is shown to be a tri-layered, level-dependent quasi-birth-and-death (LDQBD) process, and we provide a necessary and sufficient condition for the positive recurrence of LDQBDs using classical techniques. Moreover, we apply efficient numerical algorithms, designed to exploit the matrix-geometric structure of the model, to compute the approximate steady-state orbit size distribution and mean congestion and delay measures. The second case assumes that customers bring generally distributed service requirements while all other processes are identical to the first case. We show that the joint process of orbit size, environment state and server status is a level-dependent, M/G/1-type stochastic process. By employing regenerative theory, and exploiting the M/G/1-type structure, we derive a necessary and sufficient condition for stability of the system. Finally, for the exponential model, we illustrate how the main results may be used to simultaneously select arrival and service rates that minimize the mean time customers spend in orbit, subject to bound and stability constraints.

15. SUBJECT TERMS
Retrial queue, unreliable, quasi-birth-and-death process, stability

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Jeffrey P. Kharoufeh, Ph.D.
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER** *(Include area code)*
U	U	U	UU	175	(617) 373-2608; Email: J.Kharoufeh@neu.edu