CHAPTER 1

INTRODUCTION

Retrial queueing system is a special type of queueing system. Queueing theory has its origins in research by Agner Krarup Erlang, when he developed models to describe the Copenhagen telephone exchange. A. K. Erlang’s works on probability problems in telephone systems laid the groundwork for the development of queueing theory. The ideas have since seen applications in various areas including telecommunications, traffic engineering, computing and the design of factories, shops, offices and hospitals. For a detailed study on classical queueing models, one can refer to the contributions made by Allen (1978), Kleinrock (1975, 1976), Neuts (1984, 1989), Takagi (1988, 1991), Yechiali (1993), Parthasarathy (1987, 1998, 2003), Artalejo (2004), Choi et al (1995), Cox (1965), Lee HW (1994) and Falin et al (1993). A queueing system can be described as customers arriving for service, waiting for service when it is not immediately available, and having waited for service, they leave the system after being served.

In classical queueing theory, it is very often assumed that a customer who cannot get service immediately after arrival either joins the waiting line and is then served according to some queueing discipline or leaves the system for ever. However, the assumption about loss of customers who elected to leave the system is just a first order approximation to a real situation. Usually, such a customer returns to the system after some random period of time and tries to get service again. The classical queueing models do not take into account the phenomenon of retrials. For this reason, they cannot
be applied in solving a number of practically important problems. Retrial queues (or queues with returning customers, repeated attempts, etc.) have been introduced to solve this deficiency.

The following are just a few examples to explain the above discussed phenomenon.

Retail shopping queue: In a shop, a customer who finds that a queue is too long may wish to do something else and return later on with the hope that the queue dissolves. Similar behaviour may demonstrate some impatient customers who entered the waiting line but then discovered that the residual waiting time is too long.

Telephone systems: Everybody knows from experience that a telephone subscriber who obtains a busy signal repeats the call until the required connection is made. As a result, the flow of calls circulating in a telephone network consists of two parts: the flow of primary calls, which reflects the real wishes of the telephone subscribers, and the flow of repeated calls, which is a consequence of the lack of success of previous attempts. These considerations bring into focus the need of the retrial queues as a proper modelling of customer behaviour in classical telephone systems and in similar such systems.

Computer Networks: Consider a communication line with slotted time shared by several stations. The slot duration equals the transmission time of a single packet of data. If two or more stations transmit packets simultaneously, a collision takes place. As a result, all packets are destroyed and must be transmitted. The stations involved in the conflict would try to transmit it to the nearest slot, but then a collision is bound to occur. To avoid this, each station, independently of other stations, may either delay action until the next slot or retransmit the packet. In other words, each station introduces a random delay
before making the next attempt to transmit the packet. This simple description motivates the interest of the retrial feature in computer networks.

Retrial queueing models are mainly motivated by their potential for applications in telephone switching systems, telecommunication networks and computer systems. With the recent advancements in mobile communications, the significance of retrials is becoming more and more recognized. There is an extensive literature on the retrial queues. This new class of queueing system was studied by Falin (1990), Kulkarni & Liang (1997) and Artalejo (1999). For related literature, the advances in retrial queues can be found in the books of Artalejo & Gomez-Corral (2008). The overwhelming literature contributions consider the emphasis on retrial phenomenon in many practical applications.

Retrial queueing systems are characterized by the phenomenon that arrivals who find the server busy leave the service area and join the retrial group(orbit) to try again for their requests for service in random order and at random intervals. The general structure of a retrial queue is shown in Figure 1.1.

![Figure 1.1 General Structure of a Retrial Queue](image-url)
1.1 TYPES OF RETRIAL QUEUEING MODELS

Since the pioneering works published in 1950’s, retrial queues are being widely used to provide stochastic modelling of many problems which arise in telecommunication, computer networks and in daily life. In complex models of computer and communication systems, the repeated attempts are combined with a variety of queueing phenomena leading to a large number of variants. A few such variants are described as follows:

1.1.1 Batch Arrival Retrial Queues

The arrival is described as the flow of arrivals in batches. In batch arrival retrial queues, it is assumed that at every arrival epoch, a batch of k primary customers arrive with a probability of \( g_k \). If the channel is busy at the arrival epoch, the whole group joins the orbit. If the channel is free, then one of the arriving customers starts its service and the others form sources of repeated calls.

1.1.2 Retrials due to Balking and Impatience

Most queueing systems with retrials are motivated by computer and telecommunication applications, where a repeated attempt appears due to blocking in a system with limited service capacity. However, the existence of retrials can be due to other reasons. In single server system with retrials, the repeated attempts may occur due to impatience of the customers. A second possibility is provided by the consideration of mixed models with waiting line and orbit, where a customer finding a long queue upon arrival may decide to attend another secondary job and return back later, hoping to find a shorter queue.
1.1.3 Retrial Queues with Non-Persistent Customers

Let us take the case of a calling customer deciding to abandon the system after some unsuccessful retrials. This practical variant can be modelled with the help of the persistence function \( \{ H_j : j \geq 0 \} \), where \( H_j \) represents the probability that after the \( j^{\text{th}} \) attempt fails, a customer will make the \( (j+1)^{\text{th}} \) one.

1.1.4 Retrial Queues with Server Breakdowns

Most works about queueing theory assume that the service stations never fail; nevertheless, this supposition is evidently unrealistic. Indeed, service interruptions due to server breakdowns occur in practice. Since the performance of a system may be heavily affected by the service station breakdowns, such systems with repairable server are worth investigating from the queueing theory point of view as well as reliability view point. The breakdowns may be either active or passive, depending on the fact if failures occur in a busy or idle period of the server. Besides, failures can take place after a random duration of service time or just before starting the service.

1.1.5 Retrial Queues with Vacations

In several situations, the server is unavailable to the customers due to server’s failure and/or inadequate number of customers for starting a service in case of bulk service models. Those periods for which the server is unavailable is referred to as ‘server vacation periods’. The server may then perform other tasks such as maintenance work or serving secondary customers. The aim of studying the retrial queueing model with vacation is that, by utilizing the idle time (vacation period) of the server, it is possible to minimize the total average cost involved. The following paragraphs discuss certain types of vacation models.
1.1.5.1 Single vacation

At a service completion epoch, if the required number of customers is not available in the orbit to start a service, then the server leaves for a vacation of random length. After completing this vacation, upon returning, if required number of customers is available, the server begins the service. Otherwise the server will remain in the system till the orbit size reaches the required level. This type of vacation is called single vacation.

1.1.5.2 Multiple vacations

The server avails a vacation every time the system becomes empty or when the number of customers waiting for service is less than a predetermined value. After completing a vacation, if the server finds inadequate number of customers waiting again in the orbit, then the server avails of another vacation and continues to do so until it finds a sufficient number of customers to start a service. These type of vacations are called multiple vacations.

1.1.5.3 Modified vacations

After completing a vacation, if the server finds inadequate number of customers waiting in the orbit, then it begins another vacation and continues in this manner until it either finds the required number of customers or after it completes the specified number of vacations, say, ‘M’. After ‘M’ vacations, upon returning, if adequate number of customers in the orbit is available, then it will start the service, otherwise it will remain in the system till the orbit size reaches the required number. Resting on this assumption, the maximum number of vacations a server can avail is restricted to M, during an idle period. This type of vacation is called modified vacation.
1.1.5.4 Retrial queues with working vacations

The single server takes a working vacation at times when the customers being served depart from the system and no customers are in the orbit. In vacation models, during vacation times the server will not serve any customer. However, in the case of working vacation models, the arriving customers are served with a rate smaller than the regular service rate during the vacation periods. If the slow service is completed prior to the end of vacation, then the server avails of the remaining vacation. On the other hand, if the service time is extended beyond the vacation time, the server switches to regular service rate and becomes idle after service completion. The server works with different service rates rather than completely stopping service during a vacation.

1.1.6 Retrial Queues with Two Phases of Heterogeneous Service

Two phases of heterogeneous service are described as the service process in which the server provides the first essential service to all arriving customers, while some of them receive second optional service. Queueing systems in which the server provides to each customer two phases of heterogeneous service in succession have been proved very useful to model networks, production lines and telecommunication systems, where messages are processed in two stages by a single server. Retrial queueing system with two phases of heterogeneous service is the new class of queueing system in which server provides the first essential service to all arriving primary customers or the customers from the retrial group if the server is idle, whereas some of them receive second optional service.
1.1.7 Retrial Queues with Negative Arrivals

G-queues or queues with negative arrivals have been found useful to model multiprocessor computer systems, neural networks, communication systems and manufacturing settings. An arriving negative customer will vanish if the server is idle or down or on vacation. Generally, negative customers cannot accumulate in a queue and do not receive service. The arriving negative customers affect the queue behaviour mainly in the following three ways: (i) the arriving negative customer removes all the customers in the system; (ii) the arriving negative customer removes only a customer from the head of the system, including the customer in service; (iii) the arriving negative customer removes only a customer from the end of the system. The discipline in which the arriving negative customer removes only a customer from the head of the system is appropriate to model server breakdowns.

1.2 LITERATURE SURVEY

The first mathematical results about retrial queues were published in 1950’s. Since then numerous papers have been published. The growing interest of retrial queues is also reflected in the existence of a series of international workshops on retrial queues which began in Madrid (1998).

Falin & Templeton (1997) stressed fact that the standard queueing models do not take the retrial phenomenon into account and therefore, cannot be applied in solving a number of practically important problems. To emphasize this idea, they referred to the book by Kosten (1973) and Artalejo (1999). As a consequence of all these efforts, at present, the theory of retrial queues is recognized as an important part of queueing theory.
Cohen (1957) obtained steady state analytical results for the M/M/1 retrial queue. The first result on M/G/1 retrial queues is due to Kielson et al (1968), who used the method of supplementary variables. Later Aleksandrov (1974) considered the case of arbitrarily distributed service times. Choo & Conolly (1979) studied the busy period for exponentially distributed service time. Wilkinson & Radrik (1968), Wilkinson (1956) developed algorithms for numerical solutions to stationary probabilities for a retrial system with multiple servers. Fredericks & Reisner (1978) followed up their work by simplifying the balance equations provided by Wilkinson & Radnik (1968) to obtain analytical steady state results. The single server retrial queues with priority calls have been studied by Choi et al (1995, 1999, 1987). Falin (1993) and Ajmone Marson et al (2000, 2001) have studied retrial queues for many applications in telecommunication and mobile communication. Lee (1999) discussed feedback retrial queue with two types of customers.

Queueing systems with batch arrivals are common in many practical situations. In digital communication systems, messages which are transmitted could consist of a random number of packets. Falin (1976) introduced the batch arrival retrial queueing model. He used the embedded Markov chain technique to derive the joint distribution of the channel state and queue length. Another approach to the problem was proposed by Yang & Templeton (1987). Krishnakumar & Pavai Madheswari (2003) discussed some more complicated queueing situations with retrials and batch arrivals. Currently most of the retrial queueing models are studied with matrix-analytic methods. Using matrix-analytic methods, Artalejo et al (2000) and Dudin et al (1999) employ Markovian Arrival Process (MAP) as the input stream.
Queueing systems with vacation time have been found to be useful in modelling those systems in which the server has additional tasks. Many real world systems can be modelled as queues with different vacation policies. A comprehensive and excellent study on the vacation models can be found in Takagi (1988). A comprehensive survey on the recent results for a variety of vacation models can be found in the Doshi (1986, 1990), Tian (1990), Falin (1995), Krishna Reddy & Anitha (1999), Gautam Choudhury (2008), Zhao et al (2008), Xu et al (2009), and Jau - Chuan ke et al (2010). In their research work, the server operates under any one of the vacation policies: single vacation, multiple vacations, gated vacation and so on. Jau - Chuan ke et al (2009) first introduced the modified vacation policy (J vacations) to retrial queues. Arumuganathan & Jeyakumar (2005) developed multiple vacations with N- policy for a bulk queueing system with setup times and closed times. Jau – Chuan ke et al (2010) developed the variant vacation policy for an M^x/G/1 queueing system, where the server operates at N- policy and takes at most J vacations when the system is empty. Senthil Kumar & Arumuganathan (2008) analyzed a single server retrial queue with batch arrivals, two phases of heterogeneous service and multiple vacations with N-policy. Haridass & Arumuganathan (2011, 2012) analyzed a batch arrival, bulk service queueing system with interrupted vacation. A queue with working vacation was first analysed by Servi and Finn (2002), who obtained the queue length distribution of M/M/1/WV queue. The server showed a lower service rate rather than completely stopping the service during a vacation period. Subsequently, Kim et al (2003) analysed an M/G/1 queue with exponential distributed working vacations. Tien Van Do (2009) studied a Markovian retrial queue with working vacation.

Stochastic decomposition is a major result for vacation models; the pioneer work in this direction was due to Furhamann & Cooper (1985). Further, Natalia (2006) suggested the stochastic decomposition property for
the retrial queues with breakdown. The property of stochastic decomposition for the M/G/1 retrial queue was studied by many authors. The idea of using this property for deriving an explicit formula for moments of the number of customers in the orbit was suggested by Artalejo & Falin (1994).

The most classical case in a queue assumes a reliable machine or server. However, in practice, we often meet cases where the servers may fail and can be repaired. Kulkami & Choi (1990) introduced the retrial queues, considering these server failures and repairs. Queueing systems with repairable service station have been studied by many authors, Kosten (1947), Avi-Itzhak & Naor (1963), Krishnakumar et al (2002), Li & Zhao (2005), Tang (1997), Artalejo (1994) and Yue & Cao (1997). In the recent past, there has been a fast development in the literature on retrial queues. However, very few works take into account both the retrial phenomenon and unreliability of the server. For related literature, one can find the main results and methods about unreliable retrial queues in Aissani (1988, 1993, 1994), Aissani & Artalejo (1998), Atencia et al (2006, 2008), Wang et al (2001, 2008) and Yang & Li (1994).

During the last decade, there has been an increasing interest in queueing systems with negative customers due to their applications in telecommunication and computer networks. Negative arrivals are used as a control mechanism in many telecommunication and computer networks. G-queues or queues with negative arrivals were first introduced by Gelenbe (1989) for the purpose of modelling neural networks. For a comprehensive analysis of queueing systems with negative arrivals, one may refer to Gelenbe (1991, 1994, 2000), Harrison & Pitel (1993, 1996), Artalejo (2000, 2010), Do (2011) and their references. A detailed analysis of the retrial queues with negative customers subject to the server breakdowns and repairs were given by Liu et al (2009) and Wu & Yin (2011).
Uncoordinated attempt by several sources to use a single server facility can result in collision, leading to the loss of transmission and highlighting the need for retransmission. Jonin (1982), Falin & Sukharev (1985) analyzed the retrial queueing system with collision, known as the queue with double connections. Choi (1992) considered a retrial queueing model with collision arising from the specific communication protocol CSMA/CD. Gomez – Corral (2010) presented some results on the number of collisions in p-persistent CSMA/CD protocols. Krishna Kumar et al (2010) analyzed a Markovian single server feedback retrial queue with a linear retrial rate and collisions of customers.


In retrial queues, the inter-retrial times are modelled according to different disciplines depending on each particular application. In classical retrial policy, the intervals between successive repeated attempts are exponential distributed with a rate \( j \gamma \) when the orbit size is \( j \), as discussed by many authors. However, recent application to communication protocols and local area networks show that there are queueing situations in which the retrial rate is independent of the number of customers in the orbit. This constant retrial policy was introduced by Fayolle (1986), who modelled a telephone exchange system. Later, it was used for the stability of the ALOHA protocol Choi (1993) and unslotted CSMA/CD (Carrier Sense Multiple Access with Collision Detection) protocol Choi (1992) in communication systems. Artalejo & Gomez- Corral (1997) combined both policies by defining a linear retrial policy. Since Fayolle (1986), there has been a rapid
growth in retrial queues with constant retrial policy. Artalejo (1996, 2001), Choi (1993), and Gomez – Corral (1999) analysed retrial queues with constant retrial policy. This retrial policy is a useful device for modelling the retrial phenomenon in communication and computer networks where repeated attempts are made by processor units independently of the number of messages stored in each mode of the network. There are a number of papers Choi (1999), Moreno (2004), Diamond (1998) and Artalejo & Gomez- Corral (2007) devoted to algorithmic methods for retrial queues including the analysis of models with general inter arrival times.

1.3 OBJECTIVES OF THE WORK

The objective of this research is to develop analytical treatment of some retrial queueing models to obtain performance measures. This research proposes different queueing models arising in many practical situations. The proposed models have been theoretically developed and numerically justified. The main objectives of this research are:

- to design and develop a theoretical framework for various retrial queueing models
- to explore avenues for many practical applications
- to obtain certain important performance measures
- to analyze the performance measures with numerical illustration
- to discuss some interesting particular cases and special cases.
1.4 THESIS ORGANISATION

The thesis is organised as follows:

Chapter 2 deals with the analysis of single server non-Markovian retrial queue with working vacation and constant retrial policy. If the server is busy at the arrival time, the customers join the orbit to repeat their request later. If the server is idle, then the arriving customer begins its service immediately. The single server takes a working vacation at times when the customers being served depart from the system and when no customers are in the orbit. In vacation models, during the vacation, the server normally does not serve any customer. But in working vacation models, during the vacation periods the arriving customers are served at a smaller rate than the regular service rate. It is considered that if the slow service is completed prior to the end of vacation, the server avails of the remaining vacation. On the other hand, if the service time is extended beyond the vacation time, the server switches to regular service rate and becomes idle after service completion. The server works at different service rates rather than completely stopping service during a vacation. After completing a vacation, the server stays idle in the system until a customer arrives from main pool or from orbit. In this chapter, the probability generating function of number of customers in the retrial group is obtained. Various performance measures are derived. Numerical illustration is also provided.

Chapter 3 provides the analytical treatment of an $M^Y/G/1$ retrial queue with two phase service under active server breakdowns, two types of repair and multiple vacations with N-policy. It is assumed that there is no waiting space and therefore, if an arriving batch of customers finds the server idle, one of the arriving customers starts its service immediately and the rest joins a retrial group in order to seek the service again after a random amount of time. Otherwise, if the server is busy at the arrival epoch, then all the
customers join the orbit. The server provides the preliminary first essential service (FES) and may breakdown while serving customers. It is supposed that the server lifetime follows an exponential law with rate ‘v’. When the server fails, it is repaired immediately. The customer who was being served during the server failure chooses, with probability ‘q’ to enter the orbit (impatient customer) and with complementary probability ‘p’ to remain in the server for repair in order to conclude its remaining service (patient customer). Both service and repair times are assumed to have a general distribution. The repair times are general and they are different depending on the failure which occurred during the presence of patient customers or impatient customers. Therefore, a persistent customer will receive the service first and in this case, the customer has a certain priority in the service. On successful completion of first essential service, the customer may opt for second optional service (SOS) with probability α. At a first essential service completion epoch, without any request for second optional service and the orbit size being zero, the server does a secondary job (vacation) repeatedly until the number of customers in the orbit reaches the threshold value N. At a second optional service completion epoch, if the orbit size is zero the server does the secondary job (vacation) until the number of customers in the orbit reaches the threshold value N. At a secondary job completion epoch, if the orbit size is at least N, then the server remains in the system to render service for customers from main pool or from retrial group. In this chapter, the probability generating function of number of customers in the retrial group is obtained. Further, various performance measures are obtained. Particular cases are discussed. Reliability measures are discussed. The effects of several parameters on the system are also analyzed numerically.

Chapter 4 analyses an $M^p/G/1$ feedback retrial queue with non-persistent customers and multiple vacations with N-policy. It is assumed that there is no waiting space and therefore, if an arriving batch of customers finds
the server idle, one of the arriving customers starts its service immediately and the rest joins a retrial group in order to seek the service again after a random amount of time. Otherwise, if an arriving batch of customers on arrival finds the server busy, becomes impatient and leaves the system with probability \((1 - \alpha)\) and with probability \(\alpha\), they enter into the orbit (retrying pool). At a service completion epoch, if the number of customers in the orbit is zero, the server does a secondary job (vacation) repeatedly until the retrial group size reaches \(N\). At the secondary job (vacation) completion epoch, if the orbit size is at least \(N\), then the server remains in the system to render service for customers from main pool or from retrial group. At a service completion epoch, the leaving customer may either request for another service with probability \(f\) or leave the system forever with probability \(d\) (where \(f + d = 1\)). The probability generating function of number of customers in the retrial group is obtained. Various performance measures are derived. Numerical illustration is also provided.

Chapter 5 addresses an M/G/1 retrial queueing system with different classes of arrivals subject to breakdowns. The customers namely, positive, negative and partially negative arrive according to three independent Poisson processes. It is assumed that there is no waiting space, if an arriving positive customer finds the server idle, it gets the service immediately. If an arriving positive customer finds the server busy with regular service or slow service or under repair, it enters the retrial group (called orbit) to seek the service again until it finds the server ideal. If an arriving negative customer finds the server idle, it is lost. If an arriving negative customer finds the server busy with regular service or slow service, it not only removes the customer under service, but also causes the server break down. When the server fails, it is sent to repair immediately. If an arriving partially negative customer finds the server idle, it is lost. If an arriving partially negative customer finds the server busy with regular service, it slows down the server. If an arriving
partially negative customer finds the server busy with slow service, it breaks down the server. The probability generating functions of the number of customers in the retrial group as well as in the system is obtained. The effects of various parameters on the system performance are analyzed numerically. Reliability measures of this model are also discussed. Numerical illustrations are also provided.

Chapter 6 deals with the analysis of a single server retrial queue with multiple vacations and state dependent arrivals. The customers arrive according to Poisson process with different arrival rates. If the server is busy or on vacation at the arrival epoch, the customers join the orbit to repeat their request later. On the other hand if the server is idle, then the arriving customer begins its service immediately. The customers in the orbit try for service one by one when the server is idle. At a service completion epoch, if the number of customers in the orbit is zero, the server does the secondary job (vacation) repeatedly until at least one customer is found in the orbit. At a secondary job completion epoch, if there is at least one customer found in the orbit, then the server remains in the system to render service for customers from the main pool or from the retrial group. The probability generating function of number of customers in the retrial group is obtained. Various performance measures are derived. Numerical illustrations are also provided.

Chapter 7 studies the analysis of an M/G/1 retrial queue with general retrial time, modified M-vacation and collision. At the arrival epoch, if the server is idle, then the arriving customer begins its service immediately. Otherwise, at the arrival epoch if the server is busy, the arriving customer collides with the customer in service resulting in both being shifted to the orbit. After the collision, the server becomes idle. Whenever the orbit is empty, the server leaves for a vacation. At a vacation completion epoch, if the orbit size is zero, the server leaves for another vacation of the same duration.
This pattern continues until the server returns from a vacation to find at least one customer recorded in the orbit or it has already availed M vacations. If the orbit is empty by the end of the M\textsuperscript{th} vacation, the server remains idle in the system to render service for customers from main pool or from retrial group. The probability generating function of number of customers in the retrial group is obtained. Various performance measures are derived. The effective analysis of the proposed model is discussed through numerical illustrations.

In Chapter 8, an overview of all the proposed retrial queueing models and their scope for future enhancements are presented to conclude the thesis.