CHAPTER 1

Introduction

Mathematical analysis of queueing system affords a way for predicting the performance measures of a system that attempts to provide service for randomly arising demands. Since the demands for service are governed by some probability law, the theory of queues has been developed within the frame work of the theory of stochastic processes. There are many real time applications exists for queueing models such as machining and casting in production line systems, designing local area networks, process management, time sharing devices and inventory management.

"A queueing system is defined as customers arriving for service - if not immediately provided and if having waited for service - leaving the system after being served." The term "customer" is used in a general sense and does not imply necessarily a human customer. For example, a customer may indicate calls arriving at a telephone exchange or a computer program waiting for a command to run.

Queueing theory is a main branch in applied probability. Taking decisions associated with waiting lines, queueing theory is required to contribute vital information as a stochastic process by predicting some performance characteristics like average queue length, average waiting time in the queue, average idle period of the server, etc. Hence, the main objective is to attain an economic balance between the cost and the cost associated with waiting time for that service.

The origin of queueing theory was first laid by the Danish Mathematician A.K. Erlang. His pioneer work is the theory of probabilities and telephone conversations in 1909. There are many research works carried out in queueing theory by his work with real time applications.

1.1 Characteristics of Queueing System

The simple construction of a queueing system is as follows:

- Arrival pattern of customers
- Number of servers
- System capacity and
- Queue discipline

1.1.1 Arrival Pattern of Customers

Arrival pattern specifies the inter-arrival time distribution of arriving customers, and the manner in which the customers are arriving into the system. Also it indicates whether the arrival is single or bulk and its corresponding distribution. The arrivals may come from an infinite source or a finite source. There may also be different type of arrival rates depending upon the state of the system.

1.1.2 Service Pattern of Server

It is mandatory to fix the probability distribution to describe the sequence of customer service times**.** The service times may be deterministic or probabilistic. Customers may be served one by one or in batches. In batch service, the batch size may be static (fixed batch size) or dynamic (variable batch size). Service rate may vary according to the number of customers in the queue. Neuts (1967) introduced **general bulk service rule**. At a service completion epoch, if there are τ customers waiting for service, then the following rule is adhered to:

(i) For $0 \leq \tau < a$ no service

(ii) For $a \leq \tau \leq b$ serving for a batch of τ customers

(iii) For $\tau \geq b$ serving for a batch of ‘b’ customers and the remaining $\tau - b$ customers are kept waiting in the queue

1.1.3 Number of Servers

It is not possible that always the queueing system consists of one server. In many situations the number of servers is more than one and it may be finite or infinite also. In multiple server queueing models as per the nature and the requirement of the service, the servers are either arranged in parallel or series, sometimes both. The queueing system in which many servers are arranged in series is termed as multi stage queueing system.

1.1.4 System Capacity

System capacity refers to the total number of customers waiting in the queue and in the service area. It may be finite or infinite. Ultimately the system capacity provides the physical limitation of the system.

1.1.5 Queue Discipline

Queue discipline is defined as the manner in which the customers are selected for service from the queue. As per the nature law, the frequently used queue discipline is “first in first out” (FIFO). Some other commonly used queue discipline is “Last in first out” (LIFO), which is used in many situations like inventory systems, where there is no obsolescence of stored units, since the last arrived unit is easier to reach. In “service in random order” (SIRO), customers are selected randomly for service irrespective of their arrival time. In the “priority” (PRI) queues, the customers are assigned priorities upon entering into the system. The ones with higher priority are to be selected for service ahead of those with lower priority.

1.2 Kendall’s Notation

The standard notation which describes the queueing process, introduced by Kendall (1951) is given below:

$$A/B/X/Y/Z,$$

where

A characterises inter arrival time distribution of the customers,

B indicates the service time distribution,

X is the number of parallel service channels,

Y denotes the system capacity and

Z describes the queue discipline.

The Kendall notation for the queueing system with infinite capacity and queue discipline FIFO, is A/B/X. It is not mandatory to mention Y and Z.

For example, a queuing system with Poisson bulk arrival and general bulk service of minimum capacity *'a'* and maximum capacity *'b'* and queue discipline FIFO, is denoted as $M^X/\mathrm{G}(a, b)\ /1$.

1.3 Classical Queueing Models

A classical queueing system is defined as customers arriving for service, if not immediately provided and if having waited for service, leaving the system after being served. The general structure of a classical queue is shown in Fig. 1.1.

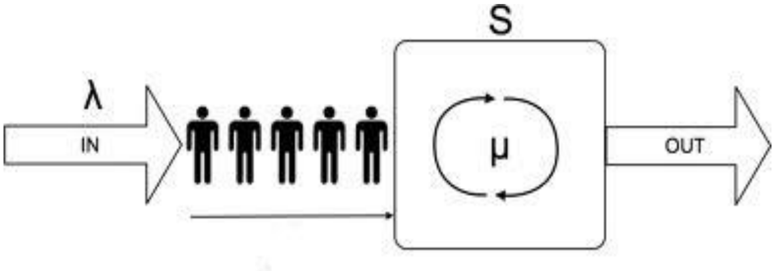

Fig. 1.1 General Structure of a Classical Queueing system

1.4 Retrial Queueing Models

An important queueing model which is used to analyse the performance characteristics of communication networks and modern computers are based on retrial queueing system. An advantage of retrial queueing system is, it considers the customer who is not receiving the service immediately (due to unavailability of server, server failure, finite capacity of the system and balking, etc.), leaves the service area and joins the retrial group (orbit), but after some random delay, come back to the system and request service again.

An orbit is a virtual queue formed by the customers, who found that the server is busy during their arrival time. The general structure of retrial queueing system is given in Fig.1.2.

In retrial queueing systems, it is essential to fix the procedure for retrials. If the probability of repeated attempts depend on the number of customers in the orbit, then it is called classical retrial policy. Also, if the probability of repeated attempts is independent of the number of orbiting customers, then it is called as constant retrial policy.

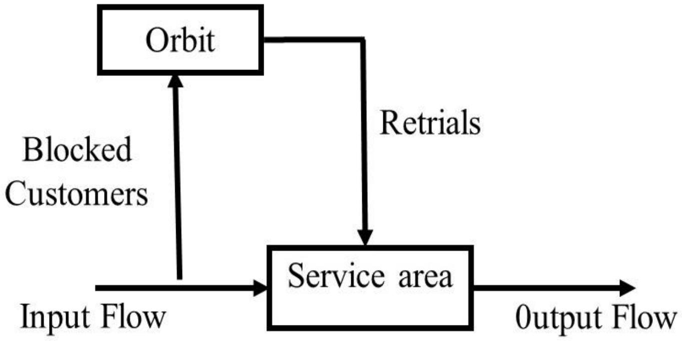

Fig. 1.2 General Structure of Retrial Queueing System

1.5 Bulk Queueing Models

If an arrival of customers into the system occur in bulk and/or service to customers done in bulk then such type of queueing model is called bulk queueing models.∑

1.6 Vacation Queueing Models

Vacation period indicates the absence of the server or non-availability of the server or idle time of the server. In some cases the server cannot start the service due to server loss and/or insufficient number of customers to start the service in case of bulk service systems. During vacation period, the server wishes to do some associated works such as preservation works or serving secondary customers. The main objective of queueing model with server vacation is effective utilization of idle time of the server. Therefore vacation queueing model tends to minimize the total average cost of the system. Some different types of vacations are given below.

1.6.1 Single Vacation

During service completion epoch, if the queue length is not adequate to start a service, then the server leaves for a vacation of random length. At the time of vacation completion, if required number of customers is available then the server starts the service, else the server will remain idle in the system until the queue length attains the threshold level. Such type of vacation is termed as single vacation.

1.6.2 Multiple Vacations

In the service completion epoch, if the queue length is not sufficient to start a service then the server leaves for a vacation. During vacation completion, if the queue length is still not sufficient to start the service, then the server takes another vacation and so on until the server finds the sufficient customers to start the service. This method of server vacation is called multiple vacations.

1.6.3 Vacation Break-off

When the server finds inadequate number of customers (in case of bulk service) waiting in the queue, then the server avails a vacation of random length. In the vacation time, if the queue length attains the threshold value to start the service, then the server breaks the vacation and switch over to primary service. The procedure of vacation break-off enables the server to start the service in normal working level even though the secondary work (vacation) is not completed.

1.6.4 Working Vacation

In vacation queueing models the server will not serve any customers at the vacation period. But in working vacation models, arriving customers during vacation period are served with lower service rate compared to regular service rate of the server. If the slow service ends earlier to the working vacation period then the server remains idle until the working vacation period ends. On the other hand if slow service exceeds the working vacation period then the slow service rate will be changed into regular service rate and becomes idle after service completion. In working vacation models the server serves in different service rates rather than completely stopping service.

1.7 Queueing Models with Server Loss or Breakdown

In many practical situations the server gets breakdown at any time while serving the customers. There exists two types of server failures in queueing models, namely server breakdown with service interruption and without service interruption.

When the server fails to provide service for the arriving customers, the service process gets interrupted and immediately it will be sent to the repair station. The service process will be continued after the repair process.

In some applications particularly in bulk service queueing systems, even when the server gets breakdown it is possible to continue or extend the service time up to some level, so that the service process will not be interrupted immediately and it will be continued for current batch of customers by doing some technical precaution arrangements. The server will be repaired after the service completion and made available for next sessions. This type of assumption will reduce the operating cost of the system which makes an impact in reducing total average cost.

1.8 Queueing Models with ‘N’ (Threshold) Policy

Threshold value ‘N’ indicates the queue length. In certain cases the server initiates service only if the queue length reaches the threshold value say ‘N’. This type of assumption in queueing model is termed as ‘N’ policy.

1.9 Queueing Models with Essential Two Phase Service

Queueing systems with two phases of service can be described as, the service process in which the same server provides two types of services called first essential service and second essential service to all arriving customers. Many research works were carried out in queueing models with two phases of service with the assumption that the first phase service is essential whereas second phase is optional. However in many real time systems two phases of service is essential to complete the service process. All the customers who have completed first essential service have to undergo second essential service in succession. Queueing systems with essential two phase services are much useful to model network, production line and telecommunication systems where messages are processed in two stages by a single server.

1.10 Queueing Models with Preparatory Work

In production line systems the server has to do some works such as alignment works or precaution arrangements before starting the service. Queueing models which adopted this feature are said to be queueing models with preparatory work.

1.11 Solution Techniques for Problems of Queueing Models

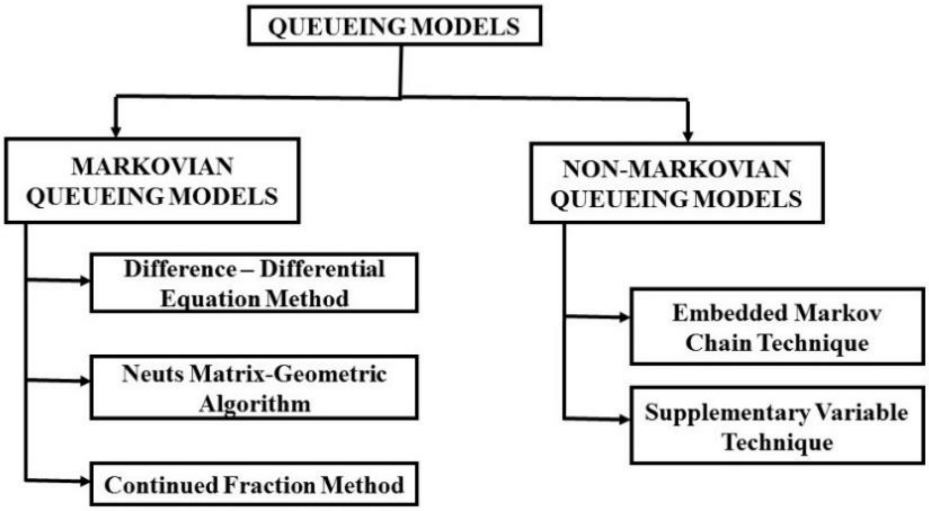
QUEUEING MODELS
MARKOVIAN QUEUEING MODELS
NON-MARKOVIAN QUEUEING MODELS
Difference – Differential Equation Method
Neuts Matrix-Geometric Algorithm
Continued Fraction Method
Embedded Markov Chain Technique
Supplementary Variable Technique

Fig. 1.3 Classification of Solution Techniques of Queueing Models

1.11.1 Markovian Queueing Models

Queueing models in which inter-arrival time and service time follows exponential distribution are called Markovian queueing models. Some of the techniques adopted to solve Markovian queueing models are given in Fig. 1.3.

1.11.1.1 Difference – Differential Equation Method

Analytical results of some queueing systems are derived by developing the corresponding difference-differential equations and solving them by using Rouche’s theorem over suitable generating functions. This method is completely discussed by Gross and Harris (1998).

1.11.1.2 Neuts Matrix-Geometric Algorithm

Neuts (1981) developed the matrix-geometric algorithmic approach to study the steady state queueing models. This method involves real arithmetic and avoids the calculation of complex roots based on Rouche's theorem.

1.11.2 Non-Markovian Queueing Models

The exponential assumption on queueing models, although very convenient, is not always realistic. There is a practical need for models that do not depend on strict Markov assumptions. Queueing models having the inter-arrival times and / or service times which are not exponentially distributed are known as non-Markovian queueing models. The techniques generally used to study non-Markovian queueing models are given in Fig. 1.3.

1.11.2.1 Embedded Markov Chain Technique

This is the first systematic technique used for solving non-Markovian queueing systems. This technique was introduced by Kendall (1953) for $M/G/1$ model. The basic idea behind this method is to simplify the description of state from two-dimensional state space to a one-dimensional state space.

1.11.2.2 Supplementary Variable Technique

This technique was introduced by Cox (1965) who introduced a supplementary variable to analyse the system $M/G/1$. By using this technique, a non-Markovian process in continuous time is converted to Markovian by the inclusion of one or more supplementary variables.

1.12 Literature Review

In this section some important literature reviews are given in the area of bulk queues, bulk queues with vacation, two phase service, server breakdown, working vacation, retrial queues, retrial queues with vacation and breakdown.

1.12.1 Classical Queueing Models

Bulk queueing models have been analysed by many researchers, which includes the study of queueing processes with bulk service (Bailey, 1954), transient analysis of bulk systems (Jaiswal, 1960). Borthakur (1971) derived steady state queue size distribution for Markovian bulk service queueing system. Medhi (1975) studied Poisson queue with a general bulk service rule and obtained the waiting time distribution under steady state condition. Neuts (1978) introduced a new technique to find steady state probabilities of a single arrival and batch service queueing system.

A complete study on bulk queueing models was given (Chaudhry and Templeton, 1984). Different models of queueing systems and its solution procedure were given (Medhi, 2002; Sheldon Ross, 2005). Krishnamoorthy and Ushakumari (2000) discussed M/M/C queueing system with accessible service batches. They computed the system size probabilities in transient as well as in steady states and also they considered arrival and departure as single whereas service is bulk.

Gupta and Goswami (2002) analysed bulk service queueing system with finite buffer in discrete-time using general bulk service rule. Benny Van Houdt studied a class of semi Markovian queues which contains many classical queues such as the GI/M/1 queue, SM/MAP/1 queue and derived queue length distribution using matrix geometric technique. Park (2011) developed a methodology to deduce the queueing and service times of individual customers whose handling cannot be observed. Maragatha Sundari et al. (2013) analysed $M^X/G/1$ Queueing system with three types of service. Performance evaluation of queueing systems by using three types of batching process has been given (Wu, 2014). Dharmaraja and Rakesh Kumar obtained transient solution of a Markovian queuing system with different servers and catastrophes. Pradhan and Gupta (2017) derived closed-form expression of bivariate probability generating function of queue content and number with the departing batch for bulk arrival and batch size dependent service queueing system using supplementary variable technique.

1.12.2 Classical Queueing Models with Vacations

Takagi (1991) explained bulk queueing systems in detail with and without vacations. Queueing systems with server vacations with several combinations was analysed by various authors. Initially Doshi (1986) made comprehensive survey of queueing systems with vacations. Lee et al. (1994) analysed an $M^X/G/1$ queueing system with 'N' policy and multiple vacations, using supplementary variable technique. A batch arrival queue with threshold was also discussed (Lee et al., 1996). Krishna Reddy and Anitha (1998) obtained stationary distribution of the number of customers in the Markovian bulk service queueing model with delayed vacations. Levy and Yechiali (1975) analysed M/G/1 queueing system with vacation. Also this model was extended by Choudhury (2002) with the inclusion of batch arrival.

Dharmaraja (2000) derived transient solution of a two processor heterogeneous system with Poisson arrival rate and exponential service times. Dharmaraja et al. (2003) presented an analytical performance model to study call blocking for non-exponential inter-arrival times. A detailed study on vacation queueing models and its applications are given (Tian and Zhang, 2006). Choudhury et al. (2007) used the concept of two phase service and derived the steady state condition for bulk arrival and single service queueing system with Bernoulli vacation schedule under multiple vacation policy.

Sikdar et al. (2008) presented some useful performance measures of interest such as probability of blocking, average queue length for finite-buffer general input queue with batch arrival and exponential multiple vacations. Strong stability of single server queue with multiple vacations is analysed (Rahmoune and Aissani, 2008). Balasubramanian et al. (2010) analysed $M^X/G(a,b)/1$ queueing system with overloading and multiple vacations. Ke et al. (2010) introduced randomized vacation policy for a batch arrival queue.

All the above discussions on queueing models are based on continuous time. Discrete time batch arrival queue with working vacation has been studied (Li et al., 2010). Wang et al. (2011) have discussed discrete time Geo/G/1 queue with randomized vacations and at most J vacations. Luo et al. (2011) discussed a recursive solution of queue length distribution for Geo/G/1 queue with vacation and different input rate.

Many authors have studied queueing models with threshold ('N'-policy). Lee and Srinivasan (1989) analysed bulk arrival queue with vacations and 'N'-policy. The different combinations of queueing systems with ‘N’-policy have been given (Lee et al., 1994, 1996). Krishnamoorthy (2002) introduced modified ‘N’-policy for $\mathrm{M/G/1}$ queues. Choudhury and Madhuchanda Paul analysed batch arrival queue with an additional service channel under ‘N’-policy. Discrete time queueing model with threshold, setup and closedown times was analysed by Pilar Moreno (2009). Ke et al. (2010) discussed a batch arrival queue with M- vacations and threshold.

Steady state analysis of non-Markovian bulk arrival and batch service queueing system with vacations and some extensions have been analysed by many researchers. Krishna Reddy et al. (1998) analysed bulk queue with ‘N’-policy multiple vacations and setup times. In their work, they have used supplementary variable technique and derived many performance measures of the system. Cost model was also given. Arumuganathan and Jeyakumar (2005) extended the above work with closedown times.

Analysis of queueing systems with control policies were also given by many authors. Arumuganathan and Ramaswami (2003) introduced state dependent arrival for bulk queueing model with multiple vacations. Jeyakumar and Arumuganathan (2011) introduced control policy on request for re-service for bulk queueing system with multiple vacations.

Markovian queueing system with working vacation and vacation interruption have been given by Li and Tian (2007). Zhang and Shi (2009) provided a study on the M/M/1 queue with Bernoulli-schedule-controlled vacation and vacation interruption. Zhang and Hou (2010) analysed non-Markovian queue with working vacations and vacation interruption. Haridass and Arumuganathan (2012) derived various performance characteristics and optimum cost analysis of $M^X/G(a,b)/1$ queueing system with vacation interruption. Extensions of vacation queueing models have been given by many authors. Vimaladevi and Ayyappan (2015) analysed controllable arrival in bulk queueing system with variant threshold for multiple vacations and setup times

1.12.3 Queueing Models with Two Phase Service

Service discipline of the queueing system involving two services, has gained lot of attention among researchers. Madan (2000) discussed a single server queue with two-stage general heterogeneous service and binomial schedule server vacations. Ke (2008) studied bulk arrival and single service queueing system with server start-up and J additional options for service. Wang et al. (2010) compared two randomized policy of M/G/1 queues with second optional service, server breakdown and start-up. Haridass and Nithya (2016) discussed two phase service for $M^X/G(a,b)/1$ queueing system with closedown times and interrupted vacation. Kuznetsov et al. (2017) analysed two-phase queueing system optimization in applications to data transmission control. Madan (2008) studied single server queue with two types of first essential service followed by two types of additional optional service and an optional deterministic server vacation.

1.12.4 Classical Queueing Models with Breakdown

Analysis of queueing system with breakdown is essential to deal with many real time systems. Shogan (1979) discussed a single server queue with arrival rate depending on server breakdown. Queueing system with time and operation dependent server failures was discussed by Shanthikumar (1982). Takine and Sengupta (1997) discussed a single server queue with service interruptions. They have characterized the queue-length distribution as well as the waiting time distribution of a single-server queue which is subject to service interruptions. Madan et al. (2003) studied steady state analysis of two $M^X/M(a,b)/1$ queueing model with random breakdowns. They considered that the repair time is exponential for one model and deterministic for another one. Wang et al. (2007) analysed the optimal control of the 'N'-policy M/G/1 queueing system with server breakdowns and general start-up times. In their system, the server is immediately turned on but is temporarily unavailable to serve the waiting customers.

Wang et al. (2009) obtained the various performance measures for the T-policy M/G/1 queue with server breakdowns and general start up times. Server breakdowns occur only when the server is in busy state and each type of breakdown requires a random number of finite stages of repair. Ke (2007) studied the operating characteristics of a batch arrival queues under vacation policies with server breakdowns and start-up/closedown times. In this paper two vacation policies namely multiple vacation and single vacation were considered. If a customer arrives during closedown times, the server is immediately started without a start-up time. Jain and Agrawal (2009) analysed the optimal policy for bulk queue with multiple types of server breakdown. An optimal control of an M /G/1 unreliable server queueing system with two phases of service and Bernoulli vacation schedule was discussed by Choudhury and Lotfi Tadj (2011).

It is clear that if breakdown occurs the server is allowed to interrupt immediately. But in most of the situations it is not possible to disturb the server before completing its batch of service. This stimulates Arumuganathan and Malliga (2006) to model bulk arrival and batch service system with breakdown without service interruption. They have described the above model as when the server got breakdown there is no need to stop the service, it will be continued for some time by doing some technical arrangements, and server will be repaired after completing the batch service. Breakdown without service interruption in a bulk arrival and batch service queueing model was also studied (Jeyakumar and senthilnathan, 2012). They modelled the queueing system with closedown time and derived probability generating function of service completion epoch, vacation completion epoch and renovation completion epoch. Nithya and Haridass (2016) derived steady state condition for controllable arrival in bulk queueing system with breakdown without service interruption and multiple vacations. Praveen Kumar Agarwal (2017) studied optimal 'N'-policy for finite queue with server breakdown and state dependent rate using recursive method.

1.12.5 Retrial Queueing Models

Retrial queues have been widely used to model telecommunication and computer networks. Therefore many research articles published in the area of retrial queues. Cohen (1957) is the first person who analysed $M/M/1$ retrial queue in steady state. A supplementary variable technique for M/G/1 retrial queue has been first used by Keilson et al. (1968) to derive joint probability distribution of number of customers in the orbit. Later Neuts and Ramalhoto (1984) studied M/G/1 retrial queue in which the server is required to search for customers.

Many survey papers have been published in the retrial queues (Yang and Templeton, 1987; Falin and Templeton, 1997). One can also refer the papers of Artalejo (1999) for developments in retrial queues. Artalejo and Lopez Herrero (2000) discussed non-Markovian retrial queue with balking. They derived the limiting distribution of the number of customers in the system by using recursive approach based on the theory of regenerative processes.

Lopez - Herrero (2002) derived detailed expansion of the probabilities of the number of customers being served in a busy period for the M/G/1 retrial queueing system. Krishna Kumar and Arivudainambi (2002) studied non-Markovian retrial queue with Bernoulli schedules and general retrial times. Artalejo and Choudhury (2004) studied M/G/1 queueing system with repeated attempts and two phases of service using embedded Markov chain method. Atencia et al. (2006) analysed non-Markovian retrial queueing system with active breakdown and the Bernoulli schedule. Retrial queueing system with second optional service in discrete time was discussed by Atencia et al. (2006). In particular Choi et al. (1993), Martin and Artalejo (1995), Artalejo (1997), Aissani (2000), Li and Zhao (2005). Falin (2010) analysed single server bulk arrival queue with retrial customers.

Jeongsim Kim et al. (2010) given moments of the queue size distribution in the MAP/G/1 retrial queue. Atencia and Moreno (2005) studied a single-server retrial queue with general retrial times and Bernoulli schedule. Wang and Li (2009) analysed a single server retrial queue with general retrial times and two-phases of service. Senthil Kumar and Arumuganathan (2010) analysed performance characteristics of single server retrial queue with general retrial time, impatient subscribers, two phases of service and Bernoulli schedule.

Wu et al. (2013) have obtained orbit size and the system size distribution for discrete-time Geo/G/1 retrial queue with preferred and impatient customers. Wang et al. (2017) have studied customer's strategic behaviour and the corresponding social maximization problem in an M/M/1 constant retrial queue with the 'N'-policy.

Falin (1976) introduced group arrivals in retrial queueing system. He modelled the system as, if arriving customers find the server is busy, then entire batch joins the orbit, but when the server is free, then one of the arriving customers starts its service and the rest join the orbit. In this article joint distribution of the system state and the queue length are derived by using embedded Markov chain technique. Yang and Templeton (1987) suggested alternative method for the above problem. Matrix analytic methods for retrial queues have been given (Dudin and Klimenok, 1999). Kim et al. (2008) have associated

Markovian arrival process in retrial queues. Avram and Gomez - Corral (2006) studied bulk service MAP/PH^{L,N}/1/N G-queues with repeated attempts.

In the overwhelming literature in retrial queues, server is capable to serve only one customer at a time. But Haridass et al. (2012) have studied bulk arrival and batch service retrial queueing system with constant retrial policy. Cost estimation was also carried out by them. Vishwa Nath Maurya (2014) studied $M^X/(\mathrm{G1}, \mathrm{G2})\ /\ 1$ retrial queueing model with second phase optional service and Bernoulli vacation schedule using PGF approach. Dudin et al. (2015) discussed single server retrial queue with group admission of customers.

1.12.6 Retrial Queueing Models with Vacations

In many vacation queueing models the server leaves for vacation only if the system size is zero. On the contrary in retrial queueing models the server remains idle in each of the service completion unless the customers are present in the system. The above vacation policy is given by Fuhrmann and Cooper (1985). Madan and Choudhury (2005) discussed a single server queue with two phases of heterogeneous service under Bernoulli schedule and a general vacation time. Zhou Wenhui (2005) analysed single-server retrial queue with FCFS orbit and Bernoulli vacation. Mohamed Boualem et al. (2007) derived stochastic inequalities for M/G/1 retrial queues with vacations and constant retrial policy.

Choudhury (2008) extensively analysed a single server queueing system with two phases of heterogeneous service and Bernoulli vacation schedule which operate under the so called linear retrial policy. Rein Nobel and Pilar Moreno (2008) studied discrete time retrial queueing system with single server.

Senthilkumar and Arumuganathan (2008a) gave probability generating function of the orbit size for $M^X/G/1$ retrial queueing model with multiple vacations, two phases of heterogeneous service and 'N'-policy. Senthilkumar and Arumuganathan (2008b) also studied $M^X/G/1$ retrial queue with two phases of heterogeneous service and general vacation time under Bernoulli schedule.

Banik (2009) studied an infinite-buffer single server queue with variant multiple vacation policy and batch Markovian arrival process. Senthil Kumar and Arumuganathan (2009) analysed M/G/1 retrial queue with non-persistent calls, two phases of heterogeneous service and different vacation policies. Murtuza Ali Abidini et al. (2016) studied retrial queueing model with vacation, 'N'-policy, polling models and the gated service.

1.12.7 Retrial Queueing Models with Breakdown

Mathematical modelling of retrial queueing systems with server breakdown is essential to analyse many real time applications. Krishna Kumar et al. (2002) analysed M/G/1 retrial queue with feedback and starting failures. Wang et al. (2008) have discussed a repairable M/G/1 retrial queue with Bernoulli vacation and two phase service. Wang and Zhao (2008) studied a discrete-time retrial queue with two failure modes.

Choudhury and Deka (2009) generalized both the classical $M^X/G/1$, retrial queue subject to random breakdown and Bernoulli admission mechanism as well as $M^X/G/1$ queue with second optional service and unreliable server. Falin (2010) used general distribution for repair times in M/G/1 retrial queue with an unreliable server. Shweta Upadhyaya (2010) has examined operating characteristics of an $M^X/G/1$ retrial queueing system under Bernoulli vacation schedule with setup times.

Ioannis Dimitriou and Christos Langaris (2010) discussed repairable queueing model with two-phase service, start-up times and retrial customers. Dmitry Efrosinin and Anastasia Winklerm (2011) introduced threshold-based recovery for unreliable retrial queueing system with a constant retrial rate. Krishna Kumar et al. (2010) have studied Markovian single server feedback retrial queue with linear retrial rate and collisions of customers using generating function technique, the joint distribution of the server state and the orbit length under steady-state is investigated. Shan Gao and Jinting Wang (2014) used both embedded Markov chain technique and the supplementary variable method for an M/G/1 retrial queue with non-persistent customers, where the server is subject to failure due to the negative arrivals. Here, the server searches for the customers in the orbit or remains idle, after completion of a service or a repair.

Tuan Phung-Duc (2015) studied asymptotic analysis for Markovian queues with two types of non-persistent retrial customers. Choudhury and Ke (2014) obtained orbit size distribution of an unreliable retrial queue with delaying repair and general retrial times under Bernoulli vacation schedule. Tseng-Chang Yen et al. (2016) investigated reliability and sensitivity analysis of the controllable repair system with warm standbys and working breakdown. Yang et al. (2016) studied an unreliable retrial queue with general repeated attempts and J optional vacations. Zayats et al. (2017) described queueing systems to model data transmission system for remote control of robot from board of ISS using public networks so called retrial queueing systems in series of space experiments. Ivan Atencia-Mc.Killop et al. (2018) analysed discrete time $\mathrm{Geo}^{[X]}/\mathrm{G}^{[X]}/1$ retrial queueing system with removal work and total renewal discipline. Chang et al. (2018) have studied an unreliable-server retrial queue with customer feedback and impatience.

1.12.8 Queueing Models with Working Vacations

Modelling queueing system with working vacation was first designed by Servi and Finn (2002). In this paper working vacation is defined as an arrival of customers during vacation will be served in different service rates instead of waiting for service until the vacation completion. Also they obtained queue length distribution of M/M/1/WV classical queueing system.

Queueing system with working vacation has been studied by many authors in recent times. In particular, Zhang and Hou (2010) derived steady state results of a service status and queue length distribution of a non-Markovian queuing system with working vacation and vacation interruption. Gao and Liu (2013) adapted Bernoulli schedule control to interrupt vacation in M/G/1 single working vacation queueing system. Jailaxmi et al. (2014) have used supplementary technique to derive performance characteristics of single server non-Markovian retrial queue with working vacation and constant retrial policy. Yang and Wu (2015) used matrix-geometric method to derive queue length distribution M/M/1 queue system with working vacation and 'N'-policy. Also they extended the model with breakdown and cost optimization.

In the above literature, working vacation queueing models were discussed only in continuous time systems, whereas Shweta Upadhyaya (2015) studied bulk arrival discrete time retrial queueing system with working vacation. This article concentrated on joint optimal values of vacation returning rate and service rate of the server during working vacation by using direct search method based on heuristic approach.

1.13 Objectives of the Work

The objective of this research is to compute analytical solutions of some classical and retrial queueing models and derive performance measures. This research study introduces various queueing models, which will be useful in designing many real time systems. Theoretically developed proposed models are given with suitable numerical illustrations. Important objectives of this research are given below:

- to model different classical and retrial queueing systems theoretically.
- to be motivated through many real time applications.
- to derive some important performance measures.
- to examine the performance measures with numerical illustrations.
- to obtain some interesting particular cases and special cases.

1.14 Organisation of the thesis

In chapter 2, **an $\boldsymbol{M^X/G(a,c,b)/1}$ queueing system with server loss in two service modes and vacation break-off** is considered. Server provides service in two service modes, as single or in bulk according to the queue length. The server provides single service, if the queue length reaches the threshold value 'a' and batch service of different batches, if the queue length reaches the value 'c'$(c > a)$ with maximum batch size of 'b' $(b > c)$. The server begins its service, only if the queue length reaches the value 'a'. After completing the service if the queue length is less than 'a' then the server leaves for vacation of random length. During secondary job (vacation) if the queue length attains the value 'a' then the server breaks the vacation and provides primary service. At a service completion epoch if the server is not reliable then the server is undergone with renewal of service station. After the repair process if the queue length is at least 'a' then the server performs its service or if the queue length is less than 'a' then the server leaves for vacation. Depending on the queue length the server may switch over from single service to bulk service or vice-versa only at service initiation epoch. For the proposed model probability generating function of queue size at an arbitrary time epoch is obtained. Various performance measures are derived with suitable numerical illustration. Cost model is also presented for the proposed model.

The aim of chapter 3 is to study on **bi-fold control strategy of a two phase bulk arrival queueing system with phase dependent breakdown and vacation break-off.** In this model service process is split into two phases called first essential service and second essential service. Here the occurrence of breakdown during first essential service and second essential service are different. When the server got failure during first essential service, service process is interrupted and sent to repair station immediately. On the contrary the server may breaks during second essential service but the service will not be interrupted, it performs continuously for current batch by doing some technical precaution arrangements. Server will be repaired after the service completion during renewal period. On the second essential service completion, if the queue length is less than 'a' then the server leaves for vacation. The server has to do preparatory work to initiate service after vacation. During vacation if the queue length reaches the value 'a' then the server breaks the vacation and performs preparatory work to start first essential service. Though the vacation period ends, if the queue length is still less than 'a' the server remains dormant (idle) until the queue length reaches the value 'a'. For this system probability generating function of the queue size will be obtained. Various performance measures are also derived with appropriate numerical solution. Additionally cost model is also presented.

The model under consideration in chapter 4 is $\boldsymbol{M^X/G/1}$ **queueing system with batch size dependent service and two patterns of working vacation.** In this chapter the server provides service in two service modes depending upon the queue length. The server provides single service if the queue length is at least 'a'. On the other hand the server provides fixed batch service if the queue length is at least'k' $(k > a)$. Batch service is provided with some fixed batch size 'k'. After completion of service if the queue length is less than 'a' then the server leaves for working vacation. During working vacation customers are served with lower service rate than the regular service rate. Service during working vacation also contains two service modes. For the proposed model probability generating function of the queue length at an arbitrary time will be obtained by using supplementary variable technique. Some performance measures will also be presented with suitable numerical illustrations.

In chapter 5, the work is based on **utilization of idle time in** $\boldsymbol{M^X/G(a,b)/1}$ **queueing system with secondary service and service interruption.** Arrival of customers into the system as bulk with Poisson arrival rate 'λ'. Server provides primary service in batches with minimum capacity 'a' and maximum capacity 'b'. Primary service will be initiated only if the queue length reaches the threshold value 'a'. In the service completion epoch if the queue length say τ is more than 'b' then the server provides service for only 'b' customers, remaining '$\tau - b$' customers have to wait in the queue for succeeding service. After completing a batch of service (primary), if the queue length is less than 'a' then the server provides single service called secondary service for '$a - 1$' customers. During primary service the server may get failure, but the service will be continued for current batch by doing some precaution arrangements. It will be renewed only after the current batch service completion. Repair of the server or proper maintenance of the server is called renewal time. On the primary service completion epoch if the server gets failure with probability 'δ' then the renewal of server station will be considered. On the other hand if there is no server failure and the queue length is greater than 'a' then the server provides service continuously with probability '$1 - \delta$'. Additionally if the queue length is less than 'a' then the server will switch over to secondary service with probability '$1 - \delta$'. During secondary service if the queue length reaches the threshold value 'a' then the server will break the single service and provides batch service. After completing secondary service if the queue length is less than one then the server leaves for single vacation. In the time of vacation completion if the queue length is still less than one then the server remains idle (dormant) until the queue length reaches the value 'a'. For the proposed model probability generating function of the queue size at an arbitrary time is obtained by using supplementary variable technique. Various performance measures are also provided with suitable numerical illustrations.

In chapter 6, **batch service retrial queueing system with server failure, threshold and multiple vacations** is considered. When bulk arrival of customers find the server is busy then entire customers will join in the orbit. On the other hand, if the server is free, then batch service will be provided according to general bulk service rule. Batch size varies from minimum of one and maximum of 'b' number of customers. Customers in the orbit seek service one by one through constant retrial policy, whenever the server is in idle state. The server may encounter failure during service. If the server fails then the renewal of service station will be considered with probability δ. If there is no server failure with probability '$1 - \delta$' in the service completion or after the renewal process and if the orbit is empty, then the server leaves for multiple vacations. The server keeps on availing vacation until the orbit size reaches the value 'N'. For this proposed queueing model, probability generating function of the orbit size will be obtained by using supplementary variable technique, various performance measures will be presented with suitable numerical illustrations. A real time application is also discussed for this system. Additionally, cost effective model is developed for this queueing model.

Section 1 in chapter 7 is devoted to the study **of bulk arrival and batch service retrial queueing system with server failure and two types of vacation.** Customers are arriving into the system in bulk as primary customers, according to Poisson arrival rate. Primary customers are served in batches under general bulk service rule with minimum of one and maximum of 'b' number of customers. If an arriving batch of customers with batch size 'τ' $(1 \leq \tau \leq b)$ finds that the server is free, then entire batch will be served immediately. On the other hand if the batch size is more than 'b' then service will be provided for only 'b' customers, and the remaining '$\tau - b$' customers will join the orbit. Additionally, if an arriving batch of customers finds that the server is busy then entire batch joins the orbit to explore service again. The customers in orbit attempt for service one by one with constant retrial rate 'γ'. In the service completion epoch if there are no customers in the orbit then the server leaves for working vacation. During non-working vacation (secondary job) period the server does not serve any customers. Moreover in working vacation period, arriving customers are served in a service rate lower than the regular service rate. If there are no customers in the orbit even at the working vacation completion then the server leaves for non-working vacation. The server may get failure while serving customers, but the service will not be interrupted and it will be continued for current service by doing some technical arrangements. For this proposed queueing model, probability generating function of the queue size at an arbitrary time epoch was obtained by using supplementary variable technique. Various performance measures were also derived with suitable numerical illustration.

In second section of chapter 7**, state dependent arrival in bulk retrial queueing system with immediate Bernoulli feedback, multiple vacations and threshold** is analysed**.** Primary customers are arriving into the system in bulk with different arrival rates λ_a and λ_b. If arriving customers find the server is busy then the entire batch will join to orbit. Customer from orbit request service one by one with constant retrial rate 'γ'. On the other hand if an arrival of customers finds the server is idle then customers will be served in batches according to general bulk service rule. Upon service completion, some of the customers may request for additional service as feedback. After service completion, customers may request service again with probability 'δ' or leave from the system with probability '$1 - \delta$'. In the service completion epoch, if the orbit size is zero then the server leaves for multiple vacations. The server continues the vacation until the orbit size reaches the value 'N' ($N > b$). At the vacation completion, if the orbit size is 'N' then the server becomes ready to provide service for customers from the main pool or from the orbit. For the designed queueing model, probability generating function of the queue size at an arbitrary time will be obtained by using supplementary variable technique. Various performance measures will be derived with suitable numerical illustration.

Finally, an overview of all the proposed bulk queueing models and their scope for future enhancements are presented to conclude the thesis.