Chapter 1

Introduction

1.1 Queuing theory

The Danish mathematician A.K. Erlang (1878-1929) is considered to be the founder of queueing theory, who presented the classic problem of determining how many circuits were needed to provide an acceptable telephone service while working in Copenhagen Telephone exchange. The first problem of queuing theory was raised in the modelling of telephone calls by Erlang (1909, 1918) where the holding times of conversations in telephone exchanges are studied. His works inspired engineers, mathematicians to deal with queuing problems using probabilistic methods. Since then queueing theory has been applied in many disciplines, such as health care and emergency planning, transportation (car, train and air traffic congestion control), stock management and production process planning, machine breakdowns and repairs, database management and computer networks and many others.

Queueing theory is a traditional mathematical framework for modeling and analyzing waiting lines and their performances. We encounter queues in almost all activities of our life, e.g., people waiting at a counter of a post office or bank, people in the waiting room of the doctor, airplanes waiting to take off, people waiting until they get connected to the call center, traffic jams, etc., are all situations whereby customers (people, airplanes, cars, etc.) queue up until they receive some kind of service (get money, being examined by the doctor, etc.). Apparently no one really wants to be in a queue especially when it is too long. However, given the fact that one has to spend enormous amount of time in queues, it is of great significance to analyze

these congestion situations using appropriate queueing models. When studying a queueing system, it is of the utmost importance to develop an appropriate model, and there several aspects have to be specified such as the frequency at which customers arrive, the number of present servers, the speed of the servers, the number of places in the waiting room, etc. Basically there are six characteristics of queuing system by means of which one can describe any queuing model, viz., the arrival process (A), the service process (B), the number of servers (C), the queue/buffer size (D), the service discipline (E), the size of the population source (F).

Kendall notation :

In 1951, Kendall (1951) proposed a mathematical abbreviation $(A/B/C/D/E/F)$ to represent any queuing model which is now rather standard throughout the queuing literature. For example, $M/M/1/FCFS/N/\infty$ denotes a queueing model with Poisson arrival, Poisson service, single server, first come first serve queue discipline, finite queue capacity of size N and infinite population size. Usually only the first three descriptors of $A/B/C/D/E/F$ are used whereas the remaining are common occurrence or their omission causes no ambiguity. If D is not mentioned the queue capacity is assumed to be infinite. If E is not specified the service discipline is considered to be first come first serve (FCFS) discipline. Therefore, $M/D/2$ implies a queueing model with Poisson arrival, deterministic service time distribution, two server, first come first serve service discipline, infinite buffer and infinite population, whereas $E_3/M/3/K$ implies queueing model with Erlang inter-arrival time distribution with 3 phase, exponential service time distribution, three server, finite buffer with buffer size K.

1.1.1 Some basic definitions :

Stochastic process :

Families of random variables which are functions of a parameter say time, are known as *stochastic process*, for example let us consider that $\{X(t), t \geqslant 0\}$, denotes the number of customers arrive during the time interval $[0,t)$ in a queuing model.

The set of possible values of a single random variable X_n or $X(t)$ of the stochastic process $\{X_n, n \geqslant 1\}$ or $\{X(t), t \geqslant 0\}$ is known as the state space of the stochastic process. Depending

on the parameter and state space, the stochastic process $\{X_n, n \geqslant 1\}$ or $\{X(t), t \geqslant 0\}$ may be classified into the following four types of processes:

(i) Discrete time discrete state space, (ii) Discrete time continuous state space, (iii) Continuous time discrete state space, (iv) Continuous time continuous state space.

In this thesis we have considered continuous time discrete state space stochastic process.

Markov process:

If $\{X(t), t \in T\}$ is a stochastic process such that, given the values $X(s)$, the values of $X(t), t > s$, does not depend on the values of $X(u), u < s$, then the process is said to be a Markov process. Mathematically, if for $t_1 < t_2 < ... < t_n < t$,

$$prob.\{a \leq X(t) \leq b \mid X(t_1) = x_1, ..., X(t_n) = x_n\} = prob.\{a \leq X(t) \leq b \mid X(t_n) = x_n\},$$

then the process $\{X(t), t \in T\}$ is a Markov process.

Markov chain:

A discrete state space Markov process is known as Markov chain. That is, the stochastic process $\{X_n, n = 0, 1, 2, ...\}$ is called a Markov chain, if, for $j, k, j_1, j_2, ... j_{n-1} \in N$

$$prob.\{X_n = k \mid X_{n-1} = j, X_{n-2} = j_1, ..., X_0 = j_{n-1}\} = prob.\{X_n = k \mid X_{n-1} = j\}.$$

Continuous time Markov chain:

A continuous time Markov process with discrete state space is called as continuous time Markov chain.

Counting process:

A stochastic process $\{X(t), t \geqslant 0\}$ with integral-valued state space (associated with counting of an event), such that as t increases the cumulative count can only increase is called a *counting process.* That is,

(i) $X(t) \geq 0$,

(ii) $X(t)$ is an integer,

(iii) If $s \leq t$ then $X(s) \leq X(t)$.

Counting process is a continuous time discrete state space stochastic process.

Poisson process:

A counting process $\{X(t), t \geqslant 0\}$, with probability distribution $\{p_n(t)\}$ (where $p_n(t) = prob.\{X(t) = n\}$), satisfying the following postulates called the *Poissionprocess*. The postulates are

I. **Independence** : $X(t+h) - X(t)$, the number of occurrences in the interval $(t, t+h)$ is independent of the number of occurrences prior to that interval.

II. **Homogeneity in time** : The probability $p_n(t)$ depends only on the length t of the interval and is independent of where this interval is situated, i.e. $p_n(t)$ gives the probability of of the number of occurrences in the interval $(t_1, t_1 + t)$ for every t.

III. **Regularity** : In an interval of infinitesimal length h, the probability of exactly one occurrence is $\lambda h + o(h)$ and that of more than one occurrences is of $o(h)$, i.e.

$$p_1(h) = \lambda h + o(h)$$

and

$$\sum_{k=2}^{\infty} p_k(h) = o(h)$$

here $o(h)$, denotes a function such that $\lim_{h \to 0} \left(\frac{o(h)}{h} \right) = 0$.

Note : The mean rate λ per unit time is known as parameter of the Poisson process.

Result 1. (Medhi (2006)) If $X(t)$, the number of occurrence of an event in time interval $[0, t)$, follows the Poisson postulates then

$$X(t) \sim P(\lambda t)$$

i.e. if $P_n(t) = Pr\{X(t) = n\}$ then

$$P_n(t) = \frac{e^{-\lambda t} (\lambda t)^n}{n!}.$$

Result 2. (Medhi (2006)) The interval $T(t)$ between two successive occurrences of an event of

a Poisson process $\{X(t) : t \geq 0\}$, having parameter λ, has a negative exponential distribution with mean $\frac{1}{\lambda}$, i.e.

$$T(t) \sim exp(\lambda)$$

Birth-death process

In birth-death process the state transitions are of only two types: "births", which increase the state variable by one and "deaths", which decrease the state by one. Birth-death processes have many applications in queueing theory, performance engineering, etc.

In queueing theory the birth-death process is the most fundamental example of a queueing model, the $M/M/1$ queue. An $M/M/1$ queue is a stochastic process whose state space is the set $\{0,1,2,3,...\}$ where the value corresponds to the number of customers in the system.

- Arrivals occur at rate λ according to a Poisson process and move the process from state i to $i+1$, $i \geq 0$.
- Service times have an exponential distribution with rate parameter μ and move the process from state i to $i-1$, $i \geq 1$.
- A single server serves customers one at a time according to a first-come, first-served discipline. When the service is complete the customer leaves the queue and the number of customers in the system reduces by one.
- The buffer is of infinite size.

Let $N(t)$ be the number present in the system at instant t, and $P_n(t) = prob.\{N(t) = n \mid N(0) = 0\}$, then $\{N(t), t \geq 0\}$is a Markov chain with denumerable state space $\{0,1,2,3,...\}$.
The Kolmogorov equations of the process

$$\begin{aligned} \frac{d}{dt}P_0(t) &= -\lambda P_0(t) + \mu P_1(t), \\ \frac{d}{dt}P_n(t) &= -(\lambda+\mu)P_n(t) + \lambda P_{n-1}(t) + \mu P_{n+1}(t), \qquad n \geq 1. \end{aligned}$$

We investigate the steady state solutions. Assume that such solution exist then

$$\lim_{t\to\infty} P_n(t) \equiv P_n \equiv Pr\{N = n\}.$$

Thus the Kolmogorov equations in steady state are

$$0 = -\lambda P_0 + \mu P_1, \tag{1.1}$$

$$0 = -(\lambda+\mu)P_n + \lambda P_{n-1} + \mu P_{n+1}, \qquad n \geq 1. \tag{1.2}$$

From (1.2) we have for $n = 1, 2, ...,$

$$\begin{aligned}
\mu P_{n+1} - \lambda P_n &= \mu P_n - \lambda P_{n-1}, \\
&= \mu P_{n-1} - \lambda P_{n-2}, (\text{putting } n = n-1) \\
&\quad \cdots \cdots \cdots \\
&= \mu P_1 - \lambda P_0, \\
&= 0, \quad From\,(1.1)
\end{aligned}$$

Therefore,

$$P_n = \left(\frac{\lambda}{\mu}\right)^n P_0 \tag{1.3}$$

Using normalization condition, i.e., $\sum_{n=0}^{\infty} P_n = 1$, we have

$$1 = \left\{1 + \sum_{n=0}^{\infty}\left(\frac{\lambda}{\mu}\right)^n\right\} P_0 \tag{1.4}$$

Hence a necessary and sufficient condition for the existence of a steady state is the convergence of the infinite series $\sum_{n=0}^{\infty}\left(\frac{\lambda}{\mu}\right)^n$, that is, $\frac{\lambda}{\mu} < 1$. Thus from (1.4) P_0 can be found and hence from (1.3) each P_n can be obtained.

Laplace Stieltjes Transforms:

Let $F(t)$ be a well-defined function of t specified for $t \geq 0$ and s be a complex number. If the following Stieltjes integral:

$$F^*(s) = \int_0^\infty e^{-st} dF(t)$$

converges on some s_0, the Stieltjes integral is called Laplace Stieltjes Transform of $F(t)$.

General properties of Laplace Stieltjes Transform

$F(t)$	$F^*(s) = \int_0^\infty e^{-st} dF(t)$
$F_1(t) + F_2(t)$	$F_1^*(s) + F_2^*(s)$
$aF(t)$	$aF^*(s)$
$F(t-a)$ $(a > 0)$	$e^{-sa}F^*(s)$
$F(at)$ $(a > 0)$	$F^*(s/a)$
$e^{-at}F(t)$ $(a > 0)$	$\frac{s}{s+a}F^*(s+a)$
$F'(t) = \frac{d}{dt}F(t)$	$s(F^*(s) - F(0))$
$tF'(t)$	$-s\frac{d}{ds}F^*(s)$
$\int_0^t F(x)dx$	$\frac{1}{s}F^*(s)$
$\lim_{t\to\infty} F(t)$	$\lim_{s\to 0^+} F^*(s)$
$\lim_{t\to 0^+} F(t)$	$\lim_{s\to\infty} F^*(s)$

In classical queueing models it is generally assumed that the customers arrive singly and are served individually, however, there are numerous real life situations where customers arrives in groups and are also served in groups. Such queues are referred as bulk queues. Since the main focus of the study of this thesis is on bulk queues, here below we will first briefly describe several terminologies related to the bulk queues, which is then followed by the de-

scription of queue with impatient customers and vacation queues.

1.2 Bulk queue

A bulk arrival bulk service queue is commonly termed as bulk queue in which arrival and service both are processed in groups or batches. In general, bulk queues are characterized by the arriving group size, serving batch size, inter arrival time distribution of successive arriving groups and the service time distribution of different batches.

1.2.1 Bulk arrival queue

In the study of queueing models when customers arrive to the system in groups/batches is termed as bulk arrival queueing model. In such queueing models, the arriving batch size may be fixed or may be random.

1.2.2 Bulk service queue

In bulk service queueing models, a single server serves a group of customers instead of a single customer only, following certain bulk service rule, usually pre-determined by the server. There are various bulk service rules are available in the literature, viz., fixed batch size bulk service rule, variable batch service rule, general bulk service (GBS) rule, versatile batch service rule, etc. and is defined as follows.

Bulk service rule with fixed batch size

In this bulk service rule, the server serves customers in batches of fixed batch size, say 'b', as follows. If at the beginning of the service server finds that the queue length is less than 'b' then it stops servicing and waits till the number goes up to 'b' and as soon as the number reaches 'b', server restarts service and immediately serves a batch of size 'b' customers. Bailey (1954) was the first to analyze bulk service queue with fixed batch size service capacity, and used this model in scheduling medical appointments.

Bulk service rule with variable server capacity

In this bulk service rule, the server serves the customers in batches of random size Y, with probability mass function $P(Y = i) = y_i\,(1 \leq i \leq b)$, to be decided at the beginning of the service, where b is the maximum serving capacity of the server. At the beginning of a service, if server finds that the queue length is less than the service capacity $Y = i\,(1 \leq i \leq b)$, the server does not remain idle until the queue length reaches to i, however, starts service with all the customers waiting in the queue at that time. That is, at the beginning of a service, if the service capacity of the server is i, with probability y_i, the server takes $min\,(i,$ the whole queue$)$ number of customers for providing service with probability y_i and the rest of the customers (if any) will wait for next round of service.

General bulk service (GBS) rule

In this bulk service rule, the server initiates service of the customers in batches only when there is a minimum number of customers, say 'a' $(a > 1)$, are present in the queue. However, if the number of customers (say r) in the queue is at least 'a' but not more than 'b' $(b > a)$ then entire group of 'r' $(a \leq r \leq b)$ customers are taken for service. However, if the number of customers in the queue is more than 'b', then first 'b' customers will be served and the rest will wait for next round of service. In special case when upper threshold limit is equal to the lower threshold limit, i.e., $a = b$, then this bulk service rule reduces to 'bulk service rule with fixed batch size'. Neuts (1967) introduced GBS rule which is also termed as '(a,b) rule'.

Versatile bulk service (VBS) rule

In this bulk service rule, the server serves customers in batches of random size Y, which is pre-decided at the service initiation epoch, with probability mass function $P(Y = i) = y_i\,(a \leq i \leq b)$, where '$a$' is the minimum and '$b$' is the maximum threshold capacity of the server. If at the beginning of a service, the serving capacity of the server is i, then the server takes $min\,(i,$ the whole queue $\geq a)$ number of customers for service with probability y_i. This bulk service rule was first introduced by Powell and Humblet (1986) with the name '(a,Y)' bulk service rule and later it was named as versatile bulk service (VBS) rule by Kim et al. (2004). In

special case when $a = 1$, then this bulk service rule reduces to bulk service rule with variable server capacity.

1.2.3 Batch size dependent bulk service rule

In this bulk service rule, the server serves customers in batches following any of the bulk service rule as discussed above. However, in this service mechanism the service time distribution of a batch of size 'r' $(a \leq i \leq b)$ changes dynamically depending on the serving batch size. To apply batch size dependent bulk service rule in a bulk service queue one must have the information about the distribution of the number of customers with the server, see e.g., Gupta and Banerjee (2011), Banerjee and Gupta (2012); Gupta and Banerjee (2011). The object of the study of batch size dependent bulk service queue is to reduce congestion in bulk service queueing models. Congestion in queueing model is generally measured in terms of long waiting time (delay), long queue and high rejection probability (for the case of finite buffer queue). Several attempts have been made to reduce congestion in a queueing system by applying batch size dependent bulk service rule, see e.g., Curry and Feldman (1985), Neuts (1987), Germs and Van Foreest (2010, 2013), Bar-Lev et al. (2013), Claeys et al. (2013a,b), Banerjee et al. (2011), Banerjee and Gupta (2012), Banerjee et al. (2013), Banerjee et al. (2014), Banerjee et al. (2015), Maity and Gupta (2015), Yu and Alfa (2015), Gupta and Pradhan (2015); Pradhan and Gupta (2017a,b); Pradhan et al. (2016a,b), etc.

1.3 Queue with impatient customers

An interesting phenomenon is observed in modern communication network systems is that, when many customers are already lined up in the queue, then a new arrival take a decision whether to join the system or not to join the system. This type of impatient behavior also occurs in our day to day life during our individual need of service from a service provider but need to be queued up. Customers usually tend to join a queue only when a short wait is expected. In some situations they first join the queue but depart after a while when further wait for service in the queue become intolerable. In literature of queueing models with impa-

tient customers we encounter two different kind of impatient behavior of the customers, viz., 'balking' and 'reneging'.

1.3.1 Balking

Balking is one of the important characterization of the impatient behavior of the customers in which an arriving customer decides not to join the queue, after observing the system, and leave the system without joining. Over the years various balking rules are presented by the researchers in literature which are characterized depending on the system congestion. From the perspective of an arriving customer, two types of balking rules are available in literature, viz., 'wait based balking' and 'system size-based balking', which are measured in terms of 'waiting time' and 'system size', respectively, and is defined as follows.

Wait based balking

If the expected waiting time, in the system, for getting service is known or informed to the arriving customers upon their arrival instant then they may decide to leave immediately if this time exceeds their maximum allowed waiting time. This balking phenomenon is termed as 'wait based balking' and is experienced in systems like call centers, where usually customers are told that how much time they will have to wait before an operator is available to answer the call.

System size based balking

When an arriving customer take the decision of leaving the system or joining the system, upon arrival, by observing the queue length of the system, then this type of balking phenomenon is termed as system size based balking. This type of balking, in practice, can be experienced at bank, doctors clinic, machine repairing center, etc.

1.3.2 Reneging

When an arriving customer initially join the queue and after spending some time in the queue get impatient due to long waiting time in the queue, and leave the system without getting the service then this behavior of customer's impatience is termed as 'reneging'. More precisely, in reneging, after joining the queue each customer wait for a certain length of time T to begin

the service, however, if service does not begin by that time the customer will get impatient and leave the queue without getting the service.

1.4 Vacation queue

A vacation queueing system is one in which server periodically becomes unavailable for a random period of time from the primary service station. The time spent away from the primary service station, by the server, is called a vacation period of the server. During this time period the server may be busy in serving some other queues or may be under maintenance (either due to routine maintenance or due to a breakdown). The vacation taken by the server may also be interpreted as the action taken to utilize the server in a secondary service center when there are no customers present at the primary service center. Thus, the server's vacations are useful for those systems where several users share a common server. This type of queueing models are frequently used as a tool to understand the congestion phenomenon in LAN (Local Area Network). Here below several vacation rules are defined that are available in literature.

1.4.1 Single vacation

In single vacation queueing model, at the end of each busy period the server examines the queue length and if he finds that the queue length is less than a threshold limit, then he will decide to go for a vacation. After returning from the vacation, if the server finds that the queue length is still less than the threshold limit, he will remain dormant till the queue length attains the threshold limit, otherwise, start servicing customers immediately.

1.4.2 Multiple vacation

In multiple vacation queueing model, at the end of each busy period the server examines the queue length and if he finds that the queue length is less than a threshold limit, then he will decide to go for a vacation. After returning from a vacation if the server does not find enough number of customers in the queue for servicing, he will leave for another vacation and the process will continue till he finds that the queue length is greater than or equal to the required

number of customers in the queue to start the service.

1.4.3 Variant of multiple vacation

In variant of multiple vacation queueing model, at the end of each busy period, the server examines the queue length and if he finds that the queue length is less than a threshold limit then he is allowed to take multiple vacations until either server finds enough number of customers in the queue to start another busy period or a maximum M number of consecutive vacations has been completed. After returning from Mth vacation the server will remain dormant till he finds that the queue length is greater than or equal to the threshold limit.

1.4.4 Multiple adaptive vacation

In multiple adaptive vacation queueing model, at the end of each busy period the server examines the queue length and if he finds that the queue length is less than a threshold limit then he is allowed to take a sequence of multiple vacations until either the server finds enough number of customers in the queue to start another busy period or a maximum $\tilde{M}$ (where $\tilde{M}$ is a random variable with a maximum possible value M) number of consecutive vacations has been completed. At the beginning of each vacation period the server will decide the value of $\tilde{M} = i$ $(1 \leq i \leq M)$ with probability m_i. After returning from $\tilde{M}$th vacation the server will become dormant till he finds that the queue length is greater than or equal to the threshold limit.

1.5 Literature survey

The thesis is mainly focused on the study of bulk service queueing models under two different situation (i) system size based balking and (ii) queue length dependent vacation (single and multiple vacation). On the light of the main topic of the thesis, this section presents literature survey on bulk service queues with impatient customer and vacation queues.

Literature survey on bulk service queue

Bulk service queue is one of the efficient tools for modeling the performance of the manufacturing system, production system, telecommunication system and transportation system, etc., and hence gained huge importance amongst researchers. One of the important applications of bulk service queueing models is found in group screening policy of blood samples (Bar-Lev et al. (2013)). The detail discussion on applications of bulk service queues in group testing is found in Abolnikov and Dukhovny (1999), Abolnikov and Dukhovny (2003), Bar-Lev et al. (2007), Claeys et al. (2010) and Bar-Lev et al. (2013). Abolnikov and Dukhovny (2003) concluded that it is preferable to use group testing instead of individual testing if the probability of contamination of an individual sample is small, they established that the use of group-testing is very cost effective in several cases, e.g., blood screening procedures (like detecting HIV, other bacteria, viruses from blood/urine samples), quality control for industrial production system, etc.

Bailey (1954) was the first to introduce the bulk service queue with fixed batch size bulk service rule and obtained steady state queue length distributions. Batch service queueing models have been studied extensively during the past decades, in continuous time setup, see e.g., Neuts (1967), Neuts (1987), Gold and Tran-Gia (1993), Chaudhry and Gupta (1999), Abolnikov and Dukhovny (1999), Abolnikov and Dukhovny (2003), Chang and Choi (2006), Bar-Lev et al. (2007), Alfa and He (2008), Bar-Lev et al. (2013), Banerjee et al. (2015), Pradhan and Gupta (2017a) etc., as well as in discrete time setup, see e.g., Chakravarthy (1993), Alfa et al. (1995), Gupta and Laxmi (2001), Gupta and Goswami (2002), Gupta and Goswami (2002), Chaudhry and Gupta (2003), Gupta et al. (2007), Claeys et al. (2010), Claeys et al. (2013b), Claeys et al. (2013a), Yu and Alfa (2015) etc. For more detail on bulk service queue we refer the readers to an excellent book by Chaudhry and Templeton (1972).

Bulk service queues with correlated arrivals have been extensively studied since last several decades. Chakravarthy (1993) studied a finite buffer $MAP/G^{(a,b)}/1/b$ queue where the maximum buffer space is considered to be equal to the maximum server capacity. Later Gupta and Laxmi (2001) analyzed $MAP/G^{(a,b)}/1/N$, in which they have considered $N > b$, and obtained the queue length distributions at various epoch, by using the supplementary variable and

the embedded Markov chain technique. Recently, Banik (2009b) analyzed $BMAP/G^{(a,b)}/1/N$ and $BMAP/MSP^{(a,b)}/1/N$ queues and obtained the queue length distributions.

It has been noticed that very few researchers have paid attention to the study of finite/infinite buffer bulk service queues in which service time distributions or service rates changes dynamically depending on the size of the batch under service or batch size dependent service rule. Neuts (1967) studied an infinite buffer $M/G^{(a,b)}/1$ queue with batch size dependent service and using the embedded Markov chain technique he obtained the queue length distribution at service completion epoch. Further, Neuts (1987) considered $M/G/1$ bulk service queue with GBS rule and using matrix analytic method obtained the joint stationary distribution of waiting time of an arriving customer and the size of the group in which customer is being served. Curry and Feldman (1985) studied the $M/M_r^{(a,b)}/1$ queue and using matrix geometric method obtained the joint distribution of the number in the queue as well as in service. Abolnikov and Dukhovny (1999, 2003) considered the batch size dependent service in bulk service queue and obtained the probability generating function (pgf) of the number of customers in the system at successive service completion epoch. Bar-Lev et al. (2007) considered $M/G_r^{(a,b)}/1$ queue and using the embedded Markov chain technique obtained an explicit expression for probability generating function of queue length distribution at departure epoch, however, due to the complexities involved in the inversion of the Laplace transform of probability generating function, they modified their model to a finite buffer queue by truncating the associated transition probability matrix at a sufficiently large value and obtained the queue length distribution at departure epoch. Using the method of roots, Chaudhry and Gai (2012) obtained the queue length distribution at departure epoch by inverting the probability generating function given in Bar-Lev et al. (2007). Recently, Bar-Lev et al. (2013) studied the two-stage group testing process of blood samples, modeled as $M/G_r^{(a,b)}/1$ queueing system. Claeys et al. (2010), Claeys et al. (2013a,b) analyzed batch service queueing model with batch size dependent service in discrete time setup. Maity and Gupta (2015) studied an infinite buffer bulk Poisson queue with versatile bulk service rule and by using matrix analytic method obtained the steady state joint probabilities of the queue content and server content. Recently, Gupta and Pradhan (2015), Pradhan et al. (2016a), Pradhan et al. (2016b), Pradhan and Gupta (2017b) considered an in-

finite buffer bulk service non-Poisson queue with batch size dependent service and obtain the steady state joint probability of server content and queue content at departure epoch using a bivariate probability generating function method. Then using supplementary variable technique they obtained a relation between the steady state joint probability of server content and queue content at service completion epoch and arbitrary epoch. Recently, Pradhan and Gupta (2017a) considered an infinite buffer $MAP/G_r^{(a,b)}/1$ queue and using bivariate vector generating function, obtained the joint distribution of queue content, server content and phase of arrivals at departure epoch and arbitrary epoch.

The finite buffer queueing models are more realistic for real world problem. Finite buffer bulk service queue with batch size dependent service has paid huge attention amongst researchers. Gupta and Banerjee (2011) considered $M/G^{(a,b)}/1/N$ queue and presented the mathematical model for obtaining joint distributions of queue content and server content at departure epoch and arbitrary epoch by using the embedded Markov chain technique and supplementary variable technique, respectively. These distributions are very important for the study of batch size dependent bulk service queues. Then they used the methodology presented in Gupta and Banerjee (2011) to study the finite buffer batch size dependent bulk service queues in Banerjee and Gupta (2012), Banerjee et al. (2013) in which they have successfully establish the fact that the batch size dependent service helps in reducing congestion in bulk service queues. Germs and Van Foreest (2010, 2013) studied bulk arrival bulk service queue with batch size as well as queue length dependent service mechanism.

Recently, Banerjee et al. (2015) considered finite buffer batch size dependent bulk service queue with *MAP* under GBS rule. By using the supplementary variable technique and the embedded Markov chain technique they obtained the joint distribution of queue content, server content and phase of the arrivals at various epochs. The finite buffer bulk service queues with batch size dependent service has been studied in discrete time setup by Banerjee et al. (2014) and Yu and Alfa (2015).

Literature survey on impatient queue

In literature of the queueing models with impatient customers we encounter two different kind of impatient behavior of the customers; 'balking' and 'reneging'. Numerous studies on queueing model with impatient phenomena of the customers have been found in literature, where server serves single customer at a time. Pioneered by Palm (1953), studies on customer's impatient behavior in a queue has been carried out by many researcher Haight (1957, 1960), Barrer (1957a,b), Ancker and Gafarian (1963a,b), TakÃ¡cs (1974), Baccelli et al. (1984), and references therein. These literature are devoted to the various balking rules depending on either the 'waiting time' and/or the 'system size' based balking. Haight (1957) studied $M/M/1$ queue with impatient customers where balking of a customer is measured in terms of constant probability. Finch (1959) considered $GI/M/1/N$ queue with system size based balking and obtained the steady state system length distribution. A more general balking rule is considered by Ancker and Gafarian (1963a,b), where the balking probabilities form a sequencing function depending on the queue length ahead of an arriving customer. Kumar et al. (1993) obtained transient solution of $M/M/1$ queue with system size based balking. The effect of system size based balking and reneging with constant reneging time has been studied by Barrer (1957a,b), Haight (1960), Daley (1965), Boots and Tijms (1999) etc., whereas in Ancker and Gafarian (1963a,b), Abou-El-Ata (1991), Abou-El-Ata and Hariri (2003), Yue et al. (2006), YUE and SUN (2008), Al-Seedy et al. (2009), etc., the reneging time of the customers are assumed to be exponentially distributed. Al-Seedy et al. (2009) obtained transient solution of the multi-server queue by considering constant balking probability and exponentially distributed reneging time. The queueing models with wait based balking have been studied by TakÃ¡cs (1974), Gavish and Schweitzer (1977), Hokstad (1979), Baccelli et al. (1984), Hu and Zazanis (1993), Perry and Asmussen (1995), Perry et al. (2000), Liu and Kulkarni (2006) etc. A comprehensive review till 2010 on the study of queuing models with impatient customer is discussed by Wang et al. (2010). The recent development in queueing models with impatient customers may be seen in the papers, Singh et al. (2014a), Chassioti et al. (2014), Laxmi and Jyothsna (2015), Goswami (2015), Saffer and Yue (2015), Guha et al. (2016) and the references therein. In those papers the impatient behavior of the customers has been studied with

$M/G/1$ or $GI/M/1$ finite or infinite queue with or without vacations. In recent years, it has been observed in literature, that there is huge interest amongst researchers to study different queueing models with equilibrium balking strategies, see e.g., Boudali and Economou (2012), Chen and Zhou (2015).

However, the balking phenomena of the joining customers has not been explored much with the bulk service queues in literature. Few researchers (Tadj et al. (1998), Jain and Pandey (2009), Laxmi and Jyothsna (2014), Wang et al. (2014), Islam et al. (2014), etc.) studied the impatient phenomenon in bulk service queue. Tadj et al. (1998) considered the 'system size based' balking with a certain threshold policy in a bulk arrival bulk service queue under 'fixed batch size' service rule, and obtained the queue length distribution in steady state. The impatient behavior of the passengers in public transport is mathematically modeled as a bulk arrival bulk service queuing system by Wang et al. (2014). They mathematically modeled the problem as general bulk service queue and obtained the mean and variance for the queue length. On public transportation problem, the situations in which passengers abandon the system after a certain amount of waiting time in a bulk arrival bulk service queuing model is analyzed by Islam et al. (2014) and studied the impact of headway variations and passenger waiting behavior on public transit performance. To the best of authors knowledge no literature has been found on bulk service queues with system size based balking which addressed to obtain the joint distributions.

Literature survey on vacation queue

The early research on classical queuing systems have addressed some complex systems where server's unavailability has been successfully used to model polling systems, maintenance models, processor failures, etc. These systems are relevant to queue with vacation. Levy and Yechiali (1975) was the first who studied the issue of efficiently utilizing server's idle time of a classical $M/G/1$ queuing system and introduced the concept of vacation queue in which the vacation time is utilized for additional work in a secondary system. They introduced the two most standard vacation policies, viz., single vacation and multiple vacation in their paper. Since the early work by Levy and Yechiali (1975), queueing theory with vacation has

been developed over the past decades, as an extension of the classical queueing theory. Since then various vacation policies, e.g. Bernoulli's vacation, variant of multiple vacation, multiple adaptive vacation, working vacation etc. have been introduced (see e.g., Doshi (1986), Takagi (1988, 1991, 1993), Tian and Zhang (2006) and so forth). In literature the vacation models may be classified on the basis of the scheduling disciplines of vacation, that is, the rule governing when the server stop servicing and take a vacation. A large number of studies, on single vacation and/or multiple vacation along with several service discipline e.g., exhaustive, limited, gated, exhaustive limited (E-limited), gated limited (G-limited), etc., are available in literature (see e.g, Shin and Pearce (1998), Krishna et al. (1998), Ho Woo Lee and Park (2001), Banik et al. (2006a), Banik et al. (2006b), Samanta et al. (2007b), Banik (2009a, 2013b); Guha and Banik (2013), Laxmi et al. (2013), Sikdar and Samanta (2016)). An excellent survey paper by Doshi (1986) summarized the major developments in this area till 1986. Finite or infinite buffer $M/G/1$ queue with vacation (single and/or multiple) has been studied by Courtois (1980), Lee (1984), Frey and Takahashi (1997) etc. Takagi (1991, 1993) provides a complete analysis of $M/G/1$ type and $Geo/G/1$ type vacation queueing systems.

The bulk service queue with vacation has been studied by Nadarajan and Subramaniyam (1984), Krishna et al. (1991), Lee et al. (1996, 1992), Krishna and Anitha (1998), Krishna et al. (1998), Krishna and Anitha (1999a), Krishna and Anitha (1999b), Gupta and Sikdar (2004a), Sikdar and Gupta (2005a), Sikdar and Gupta (2005b), Sikdar and Gupta (2008), Sikdar (2008), Sikdar and Samanta (2016), etc. Nadarajan and Subramaniyam (1984) considered $M/M^{(a,b)}/1$ queue with server vacation and using matrix geometric method obtained the queue length distributions in steady state. Krishna et al. (1991) considered the same queueing model under multiple vacation and an additional server when the queue length exceeds a preassigned number. Lee et al. (1992) analyzed the $M/M^{(1,b)}/1$ queue with single vacation and derived the expression for probability generating function of queue length distributions at departure and arbitrary epoch. Lee et al. (1996) analyzed a fixed batch size bulk service queue with single and multiple vacation and obtained probability generting function of queue length distributions at arbitrary epoch and departure epoch. Krishna and Anitha (1998) analyzed a $M/M^{(a,b)}/1$ queuing system with multiple vacation and a changeover time, and obtained a closed form

expression for the steady state queue length distribution, waiting time distribution, expected waiting time and expected queue length. Krishna and Anitha (1999a) analyzed $M/G^{(a,b)}/1$ queue with M different types of vacation policy and using the supplementary variable technique they obtained the queue length distribution and the expected queue length. Further, they discussed $M/G^{(a,b)}/1$ queue with multiple vacations in which the distribution of first vacation is different from subsequent vacations. Finite and infinite buffer $M/G^{(a,b)}/1$ queue with single vacation have been extensively studied by Gupta and Sikdar (2004a); Sikdar and Gupta (2005a). Using the supplementary variable technique they obtained the queue length distribution at various epochs. The book by Tian and Zhang (2006) provides an extensive survey on vacation queueing models till 2006.

Infinite buffer $M^X/G^{(a,b)}/1$ queueing system with vacation has been studied by Krishna et al. (1998), Arumuganathan and Jeyakumar (2005), Haridass and Arumuganathan (2011), Haridass and Arumuganathan (2012a). For recent development in the field of vacation queue the readers are referred to go through Gupta and Sikdar (2004a), Gupta and Sikdar (2004b), Sikdar and Gupta (2005b), Gupta and Sikdar (2006), Banik et al. (2006a), Sikdar (2008), Sikdar and Gupta (2008), Banik (2009a), Banik (2013a), Banik (2013b), Guha et al. (2015), Panda et al. (2016), Sikdar and Samanta (2016) for continuous time setup and Samanta et al. (2007a,b,c); Samanta and Zhang (2012) for discrete time setup, and the references therein. The vacation queueing models with correlated arrivals and bulk service have been studied by Gupta and Sikdar (2004b), Sikdar and Gupta (2005b), Sikdar (2008). Gupta and Sikdar (2004b) considered $MAP/G^{(a,b)}/1/N$ queue with single vacation and using supplementary variable technique and embedded Markov chain technique they obtained the queue length distribution at various epoch. Then Sikdar (2008) extended their research for $MAP/G^{(a,b)}/1/N$ queue with multiple vacation. Recently, Sikdar and Samanta (2016) studied $BMAP/G^Y/1/N$ queue with single and multiple vacation in an unified way and obtained the queue length distributions at various epochs. However, none of the above literature has studied bulk service vacation queues with batch size dependent service.

In most of the research on vacation queues, it has been considered that the server will go for a vacation of random length which is independent of the queue length at the vacation

initiation epoch. The vacation queueing models with queue length dependent vacation policy has been studied by Harris and Marchal (1988), Lee and Srinivasan (1989), Shin and Pearce (1998), Banik (2013a), etc. $M/G/1$ queue with queue length dependent vacation schedule and queue length dependent vacation time has been studied by Harris and Marchal (1988). They have used the well known stochastic decomposition property to obtain stationary probability distribution of queue length. Further, $M^X/G/1$ queue with E- limited service under single and multiple vacation policy with queue length dependent vacations has been studied by Lee and Srinivasan (1989). Shin and Pearce (1998) analyzed an infinite buffer $BMAP/G/1$ queuing system with Bernoulli scheduling and queue length dependent vacation policy and they derived the queue length distributions at departure epoch as well as at an arbitrary epoch. Recently, Banik (2013a) studied $BMAP/G/1/N$ queue with E-limited service and queue length dependent vacation and numerically shown that the queue length dependent vacation policy helps in reducing the congestion. To the best of authors knowledge none of the above literature has considered batch size dependent bulk service queue with queue length dependent vacation.

1.6 Objective and motivation of the thesis

The literature survey on impatient queue reveals the fact that the study of bulk service queues with impatient behavior of the arrivals, arises naturally due to congestion, received little attention from researchers. However, recent literature on bulk service queues establish successfully the fact that congestion in bulk service queues can be controlled by applying batch size dependent service policy, in which joint distribution of the queue content and server content plays an important role. This motivated us to study and analyze the bulk service queueing model with system size based balking behavior of arriving customers.

Vacation queueing model is a very effective tool to model queueing systems in which several users are using a single server. Congestion and optimization issues for queueing models with server vacation have already received attention amongst researchers. From the literature survey on vacation queuing model with bulk service, one may conclude that in past several decades many researchers successfully studied bulk service queues with several vacation poli-

cies. Recent literature on bulk service queueing models reveals that the congestion can be reduced in better way by introducing batch size dependent service policy. The queue length dependent vacation policy is also successfully reduces congestion in vacation queueing models. This motivated us to study the bulk queues with batch size dependent service and queue length dependent vacation. To analyze this type queueing models one must have the knowledge of the distribution of the number of customers in the serving batch and distribution of the queue content at vacation initiation epoch. Hence, the another objective of the thesis is to obtain the joint distribution of queue content and serving batch size, and joint distribution of the queue content and vacation type (the number of customers present in the queue at vacation initiation epoch) taken by the server for queueing models with queue length dependent single and multiple vacation. The model become more complex and appropriate for real world situation where the arrivals are bursty in nature, viz., telecommunication network, computer network, etc. This motivated us to carryout the research on batch size dependent bulk service queue with queue length dependent single vacation and multiple vacation with Markovian arrival process (MAP) and GBS rule. Hence, we can conclude that in this thesis we have successfully investigated the batch size dependent bulk service queue with queue length dependent single vacation and multiple vacation and came to the conclusion that queue length dependent vacation further reduces congestion in batch size dependent bulk service queue.

1.7 Organization of the thesis

The thesis consists of seven chapters. After giving a brief literature survey on bulk service queueing models with balking and vacation queues we have first investigated the effect of system size based balking behavior of arrivals, arises due to the congestion in bulk service queues using probability generating function method. Then we studied the batch size dependent bulk service queues with queue length dependent single vacation and multiple vacation.

In chapter 1 the general introduction about queueing theory is discussed and several terms which are used in the thesis is defined. Then a brief yet comprehensive literature survey on bulk service queues, impatient queues, and vacation queues has been presented which is followed by the objective and motivation of the thesis. At the end of this chapter the organization

of the thesis is presented.

In chapter 2, a single server Poisson queue has been considered, in which customers are served in batches of fixed size. The inter arrival times and the service times are considered to be exponentially distributed. The customers upon arrival may decide to join the system or not to join the system by observing the system length. They may join or balk the system with certain probability. Using probability generating function method we obtain the closed form expression for steady state queue length distribution, expected system (queue) length and expected waiting time of a customer in the system (queue). Finally, several numerical results are discussed in the form of table and graphs to explore the sensitivity of system parameters on key performance measures.

Chapter 3 investigates the effect of impatient phenomena of the arriving customers in a bulk service queue, where inputs are flowing into the system according to the Poisson process and are served in groups according to the 'general bulk service' rule. The service time of a group of customers follows exponential distribution. On arrival, a customer decides whether to join or balk the system, based on the observation of the system size and status of the server, i.e., whether server is busy or idle. The steady state joint probability distribution of the number of customers in the queue as well as with the server is obtained by using the probability generating function method, which is based on the roots of the characteristic equations formed using probability generating function for steady state joint probabilities. Finally, various performance measures, such as, average queue length, average waiting time, probability that the server is busy, average queue length when server is busy, etc., have been obtained. The chapter ends with several numerical discussions to demonstrate the effect of certain model parameters on the key performance measures.

Chapter 4 considers a single server finite buffer batch size dependent bulk service queue with queue length dependent vacation. Customers arrive at the system according to the Poisson process and are served in batches following the 'general bulk service rule'. The service time distribution is considered to be generally distributed which is allowed to be modulated depending on serving batch size. The server is allowed to go for a vacation, either single vacation or multiple vacations, when there is not permissible number of customers present in the queue

to start the service. The vacation time distribution is considered to be generally distributed and dynamically changes depending on the queue content at vacation initiation epoch. Using the supplementary variable technique and the embedded Markov chain technique, the joint distribution of the random variables of interest at various epochs, in steady state is obtained. Several numerical results are presented at the end to bring out the qualitative aspect of the model, which reveals the fact that queue length dependent vacation further reduces congestion in the batch size dependent bulk service queues.

Chapter 5 presents the analysis of a finite buffer bulk arrival batch size dependent bulk service queue with queue length dependent single or multiple vacations. The customers are arriving at the system in batches of random size, according to the Poisson process. A single server is providing service in batches, to the customers accepted by the system, following the 'general bulk service' rule. The server is sent for a vacation when queue content is found to be less than the minimum threshold limit of the GBS rule. Two types of vacation policies, single vacation and multiple vacations are studied in this chapter in a unified way. However, the vacation policy must be fixed at initial implementation stage and is not allowed to change in any intermediate stage. The service time distribution and the vacation time distribution are both considered to be generally distributed. The service rates are dependent on serving batch size and the vacation rates are dependent on queue length at vacation initiation epoch. The supplementary variable technique and the embedded Markov chain technique have been employed to analyze the model. Various illustrative numerical examples, to bring out the qualitative nature of the model, are presented which reveals the fact that the state dependent service rates and vacation rates eventually lead to less congestion.

Chapter 6 deals with finite buffer bulk service queue with Markovian arrival process (*MAP*) under batch size dependent service and queue length dependent single vacation or multiple vacations. A single server serves the customers in groups/batches following the 'general bulk service' rule. The service time distribution of the server is generally distributed and service rates are dependent on the number of customers within a batch under service. Whenever system becomes empty or queue length becomes less than the enough required number of customers to start the service, the server takes a vacation, either single vacation or multiple

vacations. The vacation time distribution is generally distributed and the vacation rates are dependent on the queue length at the starting point of the vacation. Using the embedded Markov chain technique and the supplementary variable technique we obtained required joint probabilities and derive the important system performance. To establish the fact that the queue length dependent vacation and batch size dependent service together will reduce congestion in the system, we have presented several numerical results at the end.

Some conclusions drawn from the research works, carried out in Chapter 2 to 6, are discussed in Chapter 7. Several future research work which may be carried out as an immediate extension of the research carried out in this thesis is also discussed.