1.1 Introduction

In a classical inventory model, servers are always available. In many practical inventory systems, servers may become unavailable for a period of time due to variety of reasons. This period of server absence may represent the server’s working on some supplementary jobs, being checked for maintenance, or simply taking a break. To analyze these systems, we introduce the server vacation in inventory models to represent the period of temporary server absence. Allowing servers to take vacation makes inventory models more realistic and flexible in the study of real-world inventory systems. The research work of vacation inventory has been applied to many practical systems such as production management, retail shop and banking system and so on. For example, cashiers in the bank, may be unavailable in the counter during the cash shortage period and the cashier may have gone to take the cash or may do some cheque/voucher clearance work. So this cashier unavailable period is called *server vacation period*. Another example is a flexible manufacturing facility is mainly used for producing customer-specified products. When there are no customers backorders, the facility switches over to produce a variety of items in stock. Due to the considerable switch over cost between “make-to-order” and “make-to-stock” the facility is not switched back to process customer orders until the number of orders is more than a critical level. Once the facility switches back to serving customer orders, the service is exhaustive. In this system, the “make-to-order” operation is a queue service process and the “make-to-stock” operation can be modeled as a server vacation.
In this thesis, we develop some inventory models with server vacation under discrete and continuous time setup. We study the effect of classical as well as new types of vacation policies on the inventory system and analyse the steady-state behaviour of the inventory models numerically. Before going to study the models, we first introduce some preliminary terminologies and the methods adopted in this thesis.

1.2 Inventory System

An inventory system represents a physical stock of items, which is being acquired or produced to meet the needs of customers. Inventory includes: raw materials, finished products, component parts, supplies, and work-in-process. An inventory system is the set of policies and controls that monitors levels of inventory and determines what levels should be maintained, when stock should be replenished, and how large orders should be. The objective of an inventory system is to minimize the total expected cost associated with the system. The research in this area focuses on the development of mathematical models of inventory systems, with the major objective being the determination of the optimal policy, i.e., a policy that minimizes the total expected cost. This theoretical research develops important insights into the inventory problems and establishes the form of the optimal inventory policies or to provide optimal decision rules. Because of a complex nature of the system, we analyse an inventory system in stochastic rather than deterministic setting, which is more appropriate to many real-life phenomena. Stochastic models are becoming increasingly important for understanding or making performance evaluation of complex systems in a broad spectrum of fields such as operations research, computer science, economics, telecommunication engineering, etc. Our aim is to have an improved understanding about the behaviour of such models, which may increase their applicability. In this thesis, we present some new stochastic models of practical interest and solution for these models.

Harris [1915] started to analyse the inventory models qualitatively and proposed the famous EOQ formula that was popularized by Wilson. Over the next few decades, numerous variations were elaborated and many of these results were collected in Raymond [1931], which was the first published book on inventory management. Arrow et al. [1951] and Arrow et al. [1958] studied a systematic account of an inventory system using renewal theoretic arguments. Dvoretzky et al. [1952a] established sufficient conditions for \((s, S)\)-policy to be optimal in the single-stage inventory problem. In the monograph of Whitin [1953] gave the relationship between the inventory management issues and clas-
sical economic thinking. Books by Gani [1957], Holt et al. [1960] and Morse [1958] were summarized a valuable review of the problems in probability theory of storage systems.

The practical applications of inventory theory were provided by Wagner [1962], Hadley and Whitin [1963] and Peterson and Silver [1979]. Numerous survey articles of stochastic inventory theory exist. The work by Veinott Jr. [1966] is still one of the most comprehensive surveys on mathematical theory of an inventory system, despite its age. The cost analysis of the different inventory systems along with the several other characteristics is given by Naddor [1966]. Gross and Harris [1971] developed the inventory system with state dependent lead time. Later Gross and Harris [1973] dealt with the idea of dependence between replenishment times and the number of outstanding orders. Aggarwal [1974], Nahmias [1978a], Wagner [1972, 1980], Silver [1981] and Porteus [1988] provided the valuable reviews of an inventory model by rigorous analysis on it. A survey of dynamic inventory problems and implementable models is given by Girlich [1984] and Hadley and Whitin [1962]. Love [1979], Bartmann and Beckmann [1992], Zipkin [2000] are few of the authors who have contributed significantly to this field. Recently, Birolini [1994] published a book on quality and reliability of some technical system and Axsäter [2006] gave the concepts on inventory control. Recently, Waters [2008] published a book in inventory control and management.

1.3 System Characteristics of Inventory Management

The inventory represents basic operations underlying a wide range of economic activities, for example, business cycle, the demand for money, the depletion of exhaustible resources and the control of production processes. A fundamental reason for holding inventory is that the demand, supply and replenishment times are uncertain and the systematic study of inventory provides an efficient way of matching them. Nahmias [1997] discussed six factors by which one can characterize an inventory system.

1. **Demand process:** One of the most important aspects of an inventory model is the customer’s demand process.

   (a) **Deterministic:** The rate of demand for units stocked by the system will be assumed to be known with certainty and to be constant.

   (b) **Stochastic:** The demand on the system cannot be predicted with certainty but instead must be described probabilistically.
2. **Lead time:** The amount of time that elapses between the placement of an order and delivery of an order. It is also a measure of the system’s response time. Often the lead time is not constant. Since the time to fill the order at the source, the shipping time and the time required to carry out the paperwork, etc., can vary from one order to another. Sometimes, the variations in the lead time are small enough so that the lead time can be assumed to be absolutely constant. However, in this work we will consider the probabilistic nature of lead time.

3. **Review policy:** Review time refers to the time points at which the inventory related aspects are obtained. The inventory literature concerns about the following types of review policies:

   (a) **Continuous review:** The inventory position is monitored continuously. As soon as the inventory position is sufficiently low/reached its specified point (called reorder point), an order is placed.

   (b) **Periodic review:** The inventory position is monitored only at equal time intervals and reorder decisions can be made only at such intervals. The intervals between these reviews are constant and the quantity ordered each time depends on the available inventory level at the time of review.

   (c) **Discrete review:** The time is measured in discrete time units with epochs $0, 1, 2, \ldots$. All events can occur only at these epochs and the system state is monitored continuously at each and every epoch.

4. **Excess demand:** Two classical situations are considered in the literature for handling the demands when the on hand inventory is zero.

   (a) **Backlog:** The demands that arrive when the inventory system is out of stock are backlogged. This means these demands are satisfied fully (*full backlog*) or partially (*partial backlog*), as soon as replenishment takes place.

   (b) **Lost Sales:** The demands arriving when the stock level is zero are completely lost. This means the demands arriving when the stock level is zero are not satisfied as soon as replenishment is received.

Recently, two other situations are considered by the researchers. They are namely

   (a) **Retrial demands:** The demands arriving when the inventory level is zero entered into the orbit. These orbiting customers compete for their demand after a random time.
(b) Postponed demands: The demands arriving when the inventory level is zero entered into the pool. These pooled customers are selected one-by-one according to some prefixed rule when the ordered items are received.

The main difference between the above two situations and the classical backlog inventory systems is that the customers have to wait even after the replenishment. In the backlog inventory models the backlogged demands are cleared (partially or completely) when the replenishment occurs. But in the retrial or the postponed demands inventory models, the orbited or pooled demands are satisfied one-by-one after the replenishment.

5. Inventory Life: Inventory goods may not be stocked for an infinite amount of time. Based on the lifetime of the item, inventory goods can be broadly classified into three categories:

(a) Deterioration: The deterioration items in the inventory system occur due to one or many factors such as storage condition, weather condition including the nature of the particular product under study. For examples of deteriorating item’s damage, spoilage, dryness, vaporization, etc.

(b) Obsolescence: Obsolescence refers to items that lose their value through time because of rapid changes of technology or the introduction of a new product by a competitor. Fashionable goods must be sharply reduced in price or otherwise disposed off after the season is over. For example, spare parts for military aircraft depend on the model and they become obsolete when a new model is introduced.

(c) Perishability: Examples of such perishable inventories include blood, fresh produced food, pharmaceutical, chemicals and assembly components in aero space and semi-conductors. While dealing with perishable systems, the consequent loss must be explicitly taken into account, as its exclusion from the model yields inaccurate performance measures. Analysis of systems stocking perishable items is far more difficult than their counterparts having an infinite lifetime and a primary reason is that the stock depletion rate is a function of the on-hand inventory level. Two main categories of life time are

i. Random life time: Items that behave independently and item that survive a random time in storage, (i.e., not fail/perish at the same time).

ii. Common life time: All items of the same age may fail/perish at the same times.
6. **Planning Horizon:** The planning horizon can be a single period, a finite number of periods, or can have infinite length.

### 1.4 Ordering Policies

The control and maintenance of inventory systems have been the integral part of any logistics system common to all sectors of the economy such as Defence and Commercial Organizations. The fundamental questions that arise in controlling the inventory are “when to order” and “how much to order”. This decision should be based on the stock, the anticipated demand and different cost factors. It is natural to think of physical stock on hand when we consider about the stock. But an ordering decision cannot be based only on the stock on hand. We must also include the outstanding orders that have not yet arrived and backorders, i.e., units that have been demanded but not yet delivered. In inventory control the stock situation is therefore characterized by the inventory position:

\[
\text{inventory position} = \text{stock on hand} + \text{outstanding orders} - \text{backorders}
\]

This means that the inventory system handles the reserved units in the same way as units that have been ordered for immediate delivery. This may be reasonable unless the delivery time is too large, which will result in unnecessary holding costs. If the delivery time is far away, then a reasonable policy is to subtract the reserved units from the inventory position first when the remaining time until delivery is less than a certain time limit. Then the customers will get their orders in time with a high probability. Although the ordering decisions are based on the inventory position, the holding cost, and shortage costs will depend on the inventory level:

\[
\text{inventory level} = \text{stock on hand} - \text{backorders}.
\]

Three most common ordering policies in connection with continuous review inventory systems are often denoted by \((R, Q)\) policy, \((s, S)\) policy and \((S - 1, S)\) policy. The periodic review inventory system has two common ordering policies, namely, \((t, S)\) policy and \((t, Q)\) policy.

1. **\((R, Q)\) policy:** When the inventory position declines to or below the reorder point \(R\), a batch quantity of size \(Q\) is ordered. If the inventory position is sufficiently low it may be necessary to order more than one batch to get above \(R\) and the considered policy is therefore, sometimes also denoted by \((R, nQ)\).
2. \((s, S)\) policy: This policy is similar to \((R, Q)\) policy. The reorder point is denoted by \(s\). When the inventory position \(i\) is declining to or below \(s\), and the quantity ordered should be \(S - i\) so that the inventory position after the order is placed is \(S\). The difference in comparison to an \((R, Q)\) policy is that, we no longer order multiples of a given batch quantity. If the demand is unit sized then both \((R, Q)\) policy and the \((s, S)\) are one and the same.

3. \((S - 1, S)\) policy: This policy simply means that we order an item whenever a demand occur, which is known as \textit{order-up-to} \(S\) or \textit{base stock policy}.

4. \((t, S)\) policy: We only can place an order at scheduling periods of length \(t\), so as to bring back the inventory level up to \(S\)

5. \((t, Q)\) policy. This policy place an order for \(Q\) items at epochs of \(t\) interval length.

The objective of the inventory control is to find an optimal ordering policy that minimizes the cost function, where the costs are associated with ordering, holding, and shortages. Specification of the system requires the demand characteristics, the order lead time characteristics, the policy for unmet demands (e.g., backlogged or lost), and the associated costs. Arrow et al. [1951] introduced the \((s, S)\) ordering policy as a way to balance the cost of the inventory with the cost of placing orders. Dvoretzky et al. [1952a] established sufficient condition on \((s, S)\) policy for single stage inventory problem to be optimal.

For a certain class of models, Scarf [1960] and Iglehart [1963] showed that a \((s, S)\) policy is optimal for fixed order cost and Veinott Jr. and Wagner [1965] gave a computational method to obtain the corresponding optimal values for \(s\) and \(S\) in the case of discrete demands. Beckmann [1964], Veinott Jr. [1966] and Porteus [1971] also provided proof for the optimality of the \((s, S)\) policy in various cost structures under the periodic review case. Gross and Harris [1971] developed continuous review \((s, S)\) inventory model with state dependent lead time. Tijms [1972] provided a detailed analysis of the inventory system under \((s, S)\) policy. Gross and Harris [1973] dealt with the idea of independence between replenishment times and the number of outstanding orders. Aggarwal [1974] presented various ordering policies for the inventory system. Continuous review \((s, S)\) inventory system had been studied by many authors in the past. Sivazlian [1974] analysed an inventory system with arbitrary inter arrival time distribution between the demands. Richards [1975] modelled an inventory with compound renewal demands. Archibald [1981] considered compound Poisson process for
the demand in the inventory system. Algorithmic approach for an \((s, S)\) inventory model was studied by Ramaswami [1981] and the optimal policy for the inventory system was calculated by Federgruen and Zipkin [1984]. Ramaswami [1981] analysed an inventory system with random lead time and two reordering levels. Kalpakam and Arivarignan [1984] discussed the \((s, S)\) policy inventory models in which demand time points formed a semi-Markovian process. An \((s, S)\) inventory system with rest periods to the server was analysed by Daniel and Ramanarayanan [1988]. Successive stories of applications of \((s, S)\) inventory models can be found in the book of Lee and Nahmias [1993]. Kalpakam and Sapna [1993] analysed an \((s, S)\) ordering policy in which items are procured on an emergency basis during stock out period. For a comprehensive review of continuous review \((s, S)\) and \((R, nQ)\) policies can be found Noblesse et al. [2014].

1.5 Queueing System

A queueing system is a system in which customers arrive to be served. A queue is formed when service requests arrive at a service facility and are forced to wait while the server is busy working on other requests. Some how, the waiting component that may occur if the server is busy seems to dominate in the description of the system, i.e. “queueing” or “waiting”. In reality, this system can also be called a service system. Although many queues can be found in everyday life (for example, at the bank teller or at the checkout counter of a supermarket), the motivation for studying queueing phenomena comes from the performance evaluation of computer and communication systems. Generally, a queueing system can be classified as four basic characteristics

- **Input process**: The input describes the pattern in which the customer arrived for service, which is generally governed by probability law. The items may arrive at the service facility in batches of fixed size or of variable size or one by one. When the case, more than one arrival is allowed to enter the system simultaneously, the input is said to be **bulk** or **batch** arriving.

- **Queue discipline**: It is a rule according to which customers are selected for service. Namely, first come first serve-FCFS (serviced in strict order of their arrival), last come first serve-LCFS (serviced in reverse order of their arrival), random order of service-ROS (serviced in random order, irrespective of their arrival), priority based service (serviced at the pre-emptive priority/non-pre-emptive priority basis).
• **Service mechanism:** It is concerned with service time and service facilities. Service time is the time interval from the start of service to the completion of service. Service facilities can be classified as a single queue with a single server, single queue with multi-server, multi-queues with one server, multi-queue with multi-server.

• **Capacity of the system:** The size of the waiting hall may be finite or infinite and the source from which customers are generated may be finite or infinite.

Hence, according to these characteristics, the customers need to use the system and minimize their delays, loss probabilities, etc. and the server wishes to minimize its cost of providing the service to customers while ensuring that the customer’s demands are satisfied.

A notation has been developed, due, for the most part to, Kendall [1953], which is now rather standard throughout the queueing literature. A queueing system is described by the series of symbols and slashes, such as $A/B/C : X/Y/Z$, so-called *Kendall’s notation*, where

- $A$ : specifies the arrival process,  
  ($M$ for a Poisson, $GI$ for a general, and $MAP$ for a Markovian arrival process),
- $B$ : specifies the service time distribution,  
  ($M$ for exponential, $PH$ for phase type and $G$ for general distribution),
- $C$ : denotes the number of servers,
- $X$ : denotes the capacity of the system,
- $Y$ : denotes the system population,
- $Z$ : specifies the queue discipline.

Since the early work by Erlang [1917] on modelling telephone traffic systems, queueing theory has received much interest by researcher over almost a century, due to its wide practical applications in many areas. Some valuable work was done by Fry [1928], Khintchine [1955], Pollaczek [1957, 1963]. A systematic treatment of the theory from the point of view of stochastic processes is due to Kendall [1951, 1953], and this has greatly influenced subsequent work in this field. Cox [1955] introduced the supplementary variable technique and Wolff [1982] named and popularized the *PASTA* (Poisson arrivals see time averages) property. Some excellent books on classical queueing theory have been published, including those by Saaty [1961], Takacs [1962], Kleinrock [1975], Cooper and Murray [1969], Cooper [1981], Wolff [1989], Gross and Harris [1998] and others.
1.6 Retrial Inventory System

Retrial queueing systems (or queueing system with repeated attempts) are characterized with the feature that a customer who finds the server busy upon arrival is obliged to leave the service area and repeat his demand after some time called retrial time. Between trials, the blocked customer joins a place of unsatisfied customers called orbit. Queues in which customers are allowed to conduct retrials have been widely used to model many practical problems in telephone switching systems, telecommunication networks and computers competing to gain service from a central processing unit. Moreover, retrial queues are also used as mathematical models for several computer systems such as packet switching networks, shared bus local area networks operating under the carrier-sense multiple access protocol and collision avoidance star local area networks, etc.

There are two major types of retrial policies are considered by the queueing theorists in the literature.

1. **Classical retrial policy:** The intervals between successive repeated attempts are exponentially distributed with parameter $n\theta$, when the orbit size is $n$.

2. **Constant retrial policy:** Retrial rate is independent of the number of units (if any) in the orbit, i.e., the retrial rate is $(1 - \delta_{0n})\theta$, where $\delta_{ij}$ denotes Kronecker’s delta. attempts are exponentially distributed random variable with parameter $\theta_n = \theta(1 - \delta_{0n}) + n\theta$. may attempt retrials from orbit.

Many of the queueing systems with repeated attempts to operate under the classical retrial policy, which was studied by Falin [1990] and Yang and Templeton [1987]. However, constant retrial policy arises naturally in a problem where the server is required to search for customers and in communication protocols of type carrier sense multiple access (CSMA). This constant retrial policy was introduced by Fayolle [1986], who investigated a telephone exchange model as an $M/M/1$ retrial queue where the customers in the retrial group form a queue and only the customer at the head of the retrial queue can request service after exponentially distributed retrial time with rate $\theta$. Farahmand [1990] calls this discipline a retrial queue with FCFS orbit retrial policy.

Later, Choi et al. [1993] generalized the constant retrial policy by considering an $M/M/1$ retrial queue with general retrial policy where only the customer at the head of the orbit may attempt retrials from orbit and retrial times (the time interval between two successive attempts made by customers) are independent and identically distributed
with general distribution. Artalejo and Gomez-Corral [1997] introduced the linear retrial policy by incorporating both classical and constant retrial policies by assuming that the time intervals between successive repeated attempts are exponentially distributed random variable with parameter \( \theta_n = \alpha (1 - \delta_n) + n\theta \), when the orbit size is \( n \). Details on retrial queues may be found in Falin and Templeton [1997], by survey papers of Falin and Artalejo [1998], Yang and Templeton [1987], Kulkarni and Liang [1997], Chakravarthy and Dudin [2003] and a recent book by Artalejo et al. [2008].

The concept of retrial demands in inventory was introduced by Artalejo et al. [2006] by assuming positive lead time and unsatisfied customer joined to the orbit and retry for their demand under the constant retrial policy. Ushakumari [2006] produced an analytical solution to a retrial inventory system with classical retrial policy. Krishnamoorthy and Jose [2007] analysed three different retrial inventory systems with positive service time and positive lead-time. Each model differs by its buffer of varying capacity, which has infinite size, the size is equal to the current level of inventory, and size is equal to the maximum inventory level \( S \) respectively. Any arriving/returning customer finds a dry inventory / server busy, with Bernoulli trails enters/returns to the orbit. For these three models, the authors assumed exponentially distributed inter-retrial times with linear retrial rate.

Sivakumar [2008] analysed a two-commodity substitutable retrial inventory system with joint ordering policy. Both the commodities are assumed to be substitutable in the sense that at the time of zero stocks of any one commodity, the other can be used to meet the demand. When the inventory positions of both commodities are zero, arriving demands enter the orbit of infinite size. Manuel et al. [2008] considered a perishable inventory system with service facilities. They assumed that demand points form a Markovian arrival process, service time is a phase type distribution, exponentially distributed life times for the items, lead time and inter-retrial time. Sivakumar and Arivarignan [2009] studied a perishable inventory system in which demands occur from a finite source under the constant retrial policy.

Lopez Herrero [2010] provided an analysis for performance measures related to waiting time and first-passage time distributions in multi-server \((s, S)\) retrial inventory models having finite retrial group. Yadavalli et al. [2011] investigated a multi-server retrial inventory system of perishable item in the stock with service facility in which apart from the usual demands, another stream of demands called negative demands, who removes one ordinary customer from the tail of the queue. They assumed that both primary and
negative demands arrive according to Markovian arrival Process and life time of each item follows an exponential distribution. Yadavalli et al. [2012] treated finite source retrial inventory system with multi-server at service facility. In both paper, lead times, service times and inter-retrial times have independent exponential distribution. Recently, Sivakumar [2011] considered single server retrial inventory system with multiple vacation, in which arrival follows Poisson and lead time, inter retrial time and vacation times have independent exponential distribution. Krishnamoorthy et al. [2012] considered a \((s, S)\) inventory model with positive service time and positive lead time, where the service process is subject to interruptions. Service interruption occurs according to a Poisson process and random interruption duration.

### 1.7 Inventory System on Postponed Demands

Postponement of service to customers takes place in different ways depending on the case of priority queues, the service to customers of lower priority stands postponed when one of the higher priority call on. In the case of pre-emptive service, customer of lower priority in service is pushed back, when a higher priority customer arrives. Deepak et al. [2004] studied a postponed demand in a queueing system. An arrival encountering the buffer full, will join the pool with Bernoulli trial. When, at a departure epoch the buffer size drops to a preassigned level \(L - 1\) \((1 < L < K)\) or below, a postponed work is transferred with probability \(p\) \((0 < p < 1)\) and positioned as the last among the waiting customers.

Krishnamoorthy and Islam [2004] analysed an \((s, S)\) inventory system with postponed demands. Here arrival of customers forms a Poisson process. When the inventory level reaches zero due to demands, further demands are sent to a pool which has capacity \(M(= \infty)\). Service to the pooled customers would be provided only after replenishment against the order placed on reaching that level \(s\)\. Further, such customers are served only if the inventory level is at least \(s + 1\). They assumed that the lead time is exponentially distributed.

Manuel et al. [2007] studied a perishable inventory system in which the positive and negative demand occurs according to independent Markovian arrival processes. Any demand that takes place when the pool (finite size \(N)\) is full, and the inventory level is zero is assumed to be lost and time interval between any two successive selections distributed as an exponential with parameter depending on the number of customers in
the pool. The waiting demands in the pool independently may renege the system after an exponentially distributed amount of time. Sivakumar and Arivarignan [2007] dealt above perishable inventory system by excluding the negative and reneging demands and assumes phase type distributed lead time. Jayaraman et al. [2012] modelled a perishable inventory system with postponed demand and multiple server vacation. They assumed Poisson demands, lead time, life times and vacation times are distributed as independent exponential distribution. Sivakumar et al. [2012] considered a discrete time postponed inventory system in which arrival occurs according to discrete Markovian arrival process and lead time follows a discrete phase type distribution.

In all the above models, the authors assumed that the pool size is finite, and the pooled demands are selected according to an exponentially distributed time lag, only when the inventory is above the reorder level. Sivakumar and Arivarignan [2009] considered a perishable inventory system in which pooled demands are selected one by one according to FCFS rule until the inventory level drops to prefixed level $N(1 \leq N \leq s)$ from the pool of infinite size. Positive and negative demands arrive according to Markovian arrival process and lead time, life time and inter-selection time are assumed to follow independent exponential distribution.

1.8 Vacation Queueing System

Queueing system with server vacation has been received considerable attention from the researchers in the last four decades. Some of the early work on queueing systems is relevant to queues with vacations are as follows: White and Christie [1958] studied queueing system with priority services and server breakdowns. Welch [1964] examined the system with exceptional service to the first customer starting a busy period. Jaiswal [1968] and Avi-Itzhak and Naor [1963] considered queues with server interruptions and different service-resumption priority rules. For more detailed study on queueing systems with server vacation, see Tian and Zhang [2006] and the references therein. A vacation policy in queueing system can be characterized by three aspects

- **Vacation start up rule:** This rule determines when the server starts a vacation. There are two major types, namely, *exhaustive* and *non-exhaustive* service. In a multi-server system, a semi exhaustive service rule may be used if some of the servers take a vacation. Another vacation start up rule is a service interruption which represents a machine failure during the operation.
• **Vacation termination/service start up rule:** This rule determines when the server resume serving the queue/system. Two popular rules are *single vacation* and *multiple vacation* policy. More general rules, such as threshold policy (*N* policy), adaptive multiple vacation policy, etc.

• **Vacation duration distribution:** Server vacations are often are assumed to be independent and identically distributed (i.i.d) random variable with any distribution, such as general distribution, exponential distribution, phase type distribution, geometric distribution.

1.8.1 **Vacation Models with Variants of Vacation Policies**

Queueing models with server vacations have been well studied in the past four decades and successfully applied in many areas such as manufacturing systems and communication network systems. Several excellent surveys of the earlier work of vacation models have been done by Doshi [1986], Takagi [1991a] and recently Ke et al. [2010b] and the book by Teghem Jr. [1986], Takagi [1991b], Tian and Zhang [2006]. The vacation queueing model has various types of vacation policies regarding the vacation start-up and vacation termination, namely

• **Vacation with exhaustive service discipline:** Server may take a vacation after served all the customer in the queue.

• **Vacation with non-exhaustive service discipline:** Server may take a vacation even when some customers are presented.
  
  – **K-limited:** After serving atmost *K* customer in the queue server goes on a vacation. Remaining customer will wait in a queue.
  
  – **Gated:** When the server arrives, customers are brought inside a gate and gate is closed, who arrives after the server starts the service will wait outside a gate. The server may take the vacation after all customers inside a gate being served. Outside a gate to be served after a vacation.
  
  – **Bernoulli:** After each service completion, the server may take a vacation with probability *p* and start a new service (if any customer present in the queue) with probability 1 − *p*. If the system is idle, after a service completion or vacation completion, the server always took a vacation (ie *p* = 1).
• **Single vacation:** Upon return from the vacation, the server finds no customer waiting in a queue, the server may wait for new customer to arrive.

• **Multiple vacation:** Upon return from the vacation, the server finds no customer waiting in a queue, the server may take another vacation until at least one customer wait in the queue.

• **J vacation:** After served exhaustively server may take at most J vacation repeatedly until at least one customer present in the queue. No customer arrives by the end of Jth vacation, the server remains dormant in the system until at least one customer arrives, at a lower rate. After completing the vacation period, the server resumes his ordinary service.

• **Synchronous vacation** (Multi-server system): All servers may take a vacation together (single or multiple) after serving exhaustively.

• **Asynchronous vacation** (Multi-server system): Some or partial servers may take (single or multiple) vacation individually and independently, if those servers are idle.

**Single Server Vacation Models**

Single server vacation models have been studied by numerous researchers since Levy and Yechiali [1975]. They studied the issue of efficiently utilizing server idle time in $M/G/1$ queue with exhaustive service discipline under single and multiple vacation. Fuhrmann and Cooper [1985] studied $M/G/1$ queue with generalized vacation under multiple vacation. They proved that the queue length distribution can be decomposed into two independent random variable, one as queue length of the ordinary queue without vacation and another one as queue length of residual vacation time. Doshi [1990] provided a systematic treatment of the exhaustive service on $M/G/1$ with exponential vacation. Takagi [1992b] analysed time dependent process on $M/G/1$ vacation queue model with exhaustive service. Frey and Takahashi [1997], Frey and Takagashi [1998] studied $M/G/1$ and $M/G/1/N$ vacation queue with exhaustive service. These studies include those by Fuhrmann [1984], Levy and Kleinrock [1986], Harris and Marchal [1988], Brill and Harris [1992], Bril and Harris [1997], Frey and Takagashi [1998] etc.

Gated and non-gated time-limited service $M/G/1$ vacation models were first studied by Leung and Eisenberg [1989, 1990]. Bischof [2001] also gave a systematic treatment of several vacation models, including gated service, limited services, decrementing service
and Bernoulli schedule service. Compared with vacation models with Poisson arrivals, the research on $GI/M/1$ vacation models started late. Tian et al. [1989] analysed $GI/M/1$ queue with multiple exponential vacation. Using matrix geometric methods, they obtained the explicit expression for a rate matrix and proved stochastic decomposition properties for the queue length and waiting time. Chatterjee and Mukherjee [1990] also studied $GI/M/1$ queue with exponential vacation. Takagi [1991a, 1993a] published a set of books that provide a complete analysis of $M/G/1$ type with both exhaustive and non-exhaustive service discipline.

The Seminal work of the Bernoulli schedule service was one by Keilson and Servi [1986]. They proposed an approximation procedure to calculate the average waiting time if an asymmetric polling system under Bernoulli schedule service discipline. The work related to Bernoulli schedule vacation model are Ramaswamy and Servi [1988]. Later, Madan [2000a,b] studied the $M/G/1$ queueing system with two phases of heterogeneous service, where the first phase of service (ordinary service) followed by a second phase of service (second optional service) respectively, under Bernoulli schedule vacation and Binomial schedule vacation. Recently, Choudhury [2008] analysed $M/G/1$ queue with the linear retrial policy and two phase service under Bernoulli vacation schedule.

An extension of the Poisson arrivals is to introduce the non-renewal arrival process in the vacation models. If the non-renewal arrival process is a Markovian arrival process, the queueing system can be treated by using the Matrix Geometric Method. $MAP/G/1$ with general vacation model was first studied by Lucantoni et al. [1990]. They matrix analytical method to establish the steady-state result on waiting time and queue length distribution. The general theory of the non-renewal arrival processes can be found in Neuts [1979] and Latouche and Ramaswami [1999]. Further, they generalize the known factorizations on queue length and waiting time for $GI/M/1$ of Doshi [1985] to $MAP/G/1$ queue. Lee et al. [2001] investigated decompositions of the queue length distributions in the $MAP/G/1$ queue with multiple and single vacation with $N$-policy Chakravarthy [2013] analysed $MAP/PH_1, PH_2/1$ queue with multiple vacations and optional secondary services. Tian and Zhang [2003a] discussed $GI/M/1$ queue with $PH$ vacation or setup vacation. Ke [2005, 2008], Ke et al. [2010a] and Chang and Ke [2009] are discussed $J$ vacation in batch arrival vacation queueing system.
Multi-server Vacation Models

In many practical queueing systems, the multi-server vacation models are more complex to analyse. Levy and Yechiali [1976] studied $M/M/c$ queue with exponential vacation and obtained distribution of a number of busy servers and the expected number of customers in the system. Also and Vinod [1986] and Kao and Narayanan [1991] discussed $M/M/c$ queue with multiple vacation for of the server, using matrix geometric approach. Provides a powerful tool in studying complex stochastic systems. Using the Matrix analytic method of Neuts [1981] and numerical method, Chao and Zhao [1998] analysed vacation model with $M/M/c$ and $GI/M/c$ type station vacation (synchronous) and server vacation (asynchronous) with exponentially distributed vacation. An $M/M/c$ queue with multiple vacation and 1-limited service can be discussed by Tyagi et al. [2002]. Recently, Krishna Kumar and Pavai Madheswari [2005] discussed $M/M/2$ queue with heterogeneous server and multiple vacation and Krishna Kumar et al. [2008] studied $MAP/PH_1, PH_2/2$ queue with Bernoulli vacations. Chakravarthy [2007] analysed the multi server queueing system with synchronous vacation. Madan et al. [2003] presented an analysis of two server vacation model with Bernoulli scheduled and a single vacation. Tian and Li [2000] and Tian and Zhang [2003b] analysed phase type distributed vacation in $M/M/c$ and $GI/M/c$ queue respectively. Zhang and Tian [2003a,b] studied multi-server queue with some server vacation of both synchronous and asynchronous type of single and multiple vacation.

Discrete Time Vacation Models

Compared with research on continuous time vacation models, there are fewer studies on discrete time vacation models. $Geo/G/1$ queue with multiple adaptive vacation was analysed by Zhang and Tian [2001]. Ishizaku et al. [1995] presented the discrete time gated service $Geo/G/1$ vacation models. In Takagi [1993b] gave a detailed analysis of finite and infinite buffer $Geo/G/1$ type queues with different vacation policies. Later, Li and Yang [1998] considered $Geo/G/1$ retrial queue with Bernoulli schedule and Zhang and Tian [2001] added multiple adaptive vacation to this model. Wang et al. [2011] considered randomized and $J$ vacations in $Geo/G/1$ queue. A discrete time $GI/Geo/1$ queue with server vacation is presented by Tian and Zhang [2002] and along with server vacation, vacation interruption was studied by Li and Tian [2007]. Discrete time $MAP$ and $PH$ distribution was presented by Neuts [1979], Alfa and Chakravarthy [1994], Alfa et al. [1995]. In Alfa [1995], Alfa and Frigui [1996], Alfa [1998a,b] studied discrete
time $MAP/PH/1$ queues respectively, with exhaustive time-limited service, $NT$-policy, priority server and gated time-limited service. Later Alfa [2003] presented a unified framework for analysing the different class of vacation models in discrete time, namely, standard vacation systems, time-limited system, number limited system and random time limited systems. Recently Alfa [2010] have a detailed study on discrete time modelling of a single node system.

1.9 Inventory System with Server Vacation

An inventory system with server vacation has received very little attention in the literature. Daniel and Ramanarayanan [1987] introduced the concept of server vacation in a continuous review inventory system with two servers. Daniel and Ramanarayanan [1988] studied an $(s, S)$ inventory system in which server takes a rest when the level of inventory is zero. The demand that occurred during stock out period were lost. The inter-occurrence times between two successive demands, the lead times, and the rest times are assumed to follow general distributions, which are mutually independent. Using renewal and convolution techniques, they obtained the state transition probabilities.

Narayanan et al. [2008] considered an inventory system with random positive service time. The server took multiple vacations whenever there was no customer waiting in the system or the inventory level was zero. Customers arrived at the service station according to a Markovian arrival process and the service time for each customer had a phase-type distribution. They assumed correlated lead time for the orders and an infinite waiting hall for the customers. The customers who wait for service may renge after a random time. Under the above assumptions, they analysed the level dependent quasi birth-death process.

Sivakumar [2011] studied a retrial inventory system with server vacation. The server took multiple vacations whenever the inventory level was zero. During the stock out period or server vacation period, arriving demands enter to orbit, orbit size is infinite. He assumed independent exponential distribution for inter-demand times, lead times, inter-retrial times and server vacation times. He illustrated some numerical example, using steady state probability vector.

Jayaraman et al. [2012] investigated the perishable inventory system with postponed demands and server vacation. The server adopts multiple vacation, whenever there was no inventory in the system. Any arriving demands find the server is on vacation/
zero inventory, enter the pool with a certain probability and leave the system with complementary probability. They assumed Poisson demands and life times of each item, lead times, server vacation time and inter-selection time between any two consecutive selections are independently distributed as exponential.

In a production inventory system, Krishnamoorthy and Narayanan [2011] developed the concept of server vacation at service facility. The server takes multiple vacation in the absence of customers or inventory or both. They assume that the production process to follow a Markovian production process (MPP) and customer arrival process follows a MAP. When inventory falls to $s$, the production process is switched on, and is switched off when the on-hand inventory reaches $S$. Duration of each vacation and service time required for each customer are assumed to be independent and non-identical phase-type distributed random variables. Under the condition of stability, they investigate the system state distribution and computed some several performance measures.

1.10 Discrete Time Inventory System

Discrete time analysis of queueing systems has become very important over the last few years with the advent of new technologies. Nowadays, telecommunication systems tend to be more of digital form than analog. For this reason, discrete time systems are more appropriate than their continuous time counterparts to model computer and telecommunications systems. In fact, the discrete time scale reveals the nature of an underlying application, e.g., the clock time unit in a computer system, the fixed size data units (bits, bytes, fixed length packets) on a communication channel, etc. Most queueing models in the literature before the early 1990s were developed in continuous time. Only the models that were based on the embedded Markov chain studied the systems in discrete time models such as the $M/G/1$ and $GI/M/1$ which were well studied in the 1950s (See Kendall [1951], Kendall [1953]). The discrete time models developed before then were few and far between. The earlier discrete time models were those by Galliher and Wheeler [1958], Dafermos and Neuts [1971], Klimko and Neuts [1973], Minh [1978]. Takagi [1993b] gave the analysis of various kinds of Geo/$G/1$ queues. Alfa [2002] and Alfa and Xue [2007] presented a computationally more efficient method for discrete time queues. More recently, Alfa [2010] published a book on discrete time queueing systems for telecommunications. He gave some fruitful methodologies and references in the book.

First, let us define discrete time analysis. Any analysis in which the system is
observed, only at specific points in time is a discrete time system. For examples in the cases where a system is observed only at points of event occurrences such as at arrivals or departures, at pre specified points which may be of equal or unequal intervals, etc. For simplicity, without loss of generality, we only consider cases where the time intervals are equal and numbered sequentially as 0, 1, 2, 3, . . . . For these presentations, arrivals and service commencements and completions which occur between time epochs n and n + 1 is assumed to occur at time n + 1, see Dafermos and Neuts [1971]. Service times are at least one unit of time longer. Even though more than one different event may occur during an interval, a rule is established in advance, to resolve conflicts, when there is one. Such rules come to play mainly at the boundaries. In dealing with such conflicts Hunter [1983] considers n− and n+ and then defines discrete system based on these. One conflict resolution approach allows arrivals at n+ and service completion at n− (Late Arrival System), and another one has been just the reversed (Early Arrival System). Gravey and Hebuterne [1992] considered departure first and also arrival first. All these are simply rules for dealing with conflict. When a queueing system has more activities than just arrivals and departures, such as having vacations and other interruptions, both the definitions of Hunter [1983] and Gravey and Hebuterne [1992] will have to be expanded accordingly.

The first paper on discrete time inventory models was by Bar-Lev and Perry [1989], who analysed a Markovian inventory model for perishable commodities, in which the arrivals of items into the system as well as the demands for these items were assumed to be discrete random variables. Each item is classified into exactly one of the N age category. New item arriving into the system are placed into the first age category. The items of age j(j = 1, 2, . . . , N − 1), which has not been removed by a demand during a period, are transferred into age category j + 1 at the beginning of the next period. The items of age N, which are not removed by demand are lost. If the demand is larger than the total number of items in the system, then the excess demand is registered as unsatisfied demand. Otherwise, the demand is satisfied according to a FIFO issuing policy.

Lian and Liu [1999] developed a discrete time inventory model with geometric inter-demand times and constant life time. They assumed that the demands arrive in batches and that the batch-size was random. They also assumed (0, S) ordering policy, with instantaneous supply, which clears any backlog and restores the stock to its maximum capacity S. They used matrix-analytic methods to construct a discrete time Markov
chain for the inventory level and they obtained a closed-form average cost function. A discrete time model for common life time inventory systems were analysed in Lian et al. [2005]. They assumed that the demand arrives in batches according to a discrete phase-type renewal process and the lifetime of an item had a discrete $PH$-distribution. They assumed that zero lead time and unsatisfied demands were completely backlogged.

Abboud [2001] analysed a discrete time Markov model for production inventory systems with machine breakdowns. He assumed that the demand and production rates were constant and the production rate was greater than the demand rate. The failure time and the repair times were independently distributed as geometric and the demands that occur during stock-out were backlogged.

Recently, a discrete-time inventory model in which demands arrive according to a discrete Markovian arrival process was treated in Sivakumar et al. [2012]. They assumed a $(s, S)$ ordering policy with discrete phase-type lead time. The demands that occur during stock-out periods either enter a pool which has a finite capacity $N(<\infty)$ or leaves the system with a predefined probability. Demands are considered to be lost, when the pool is full and the inventory level is zero. The demands in the pool are selected one by one, if the on-hand inventory level is above $s + 1$, and the interval time between any two successive selections is assumed to have discrete phase-type distribution. For more detailed on DMAP can be found in Alfa [1998b], Alfa and Frigui [1996], Chaudhry et al. [2002]. Shophia Lawrence et al. [2013] investigated a discrete time $(s, S)$ inventory system with Bernoulli demand and geometric lead time. Perishable items are packaged and stored in a suitable and controlled environment, so that item in the packets doesn’t perish over the time. However, when a packet opened to meet a unit demand, they assumed that all the remaining items of that packet may perish at times.

### 1.11 Brief Outline of the Thesis

In this thesis, we made an attempt to provide a comprehensive analysis of some continuous review $(s, S)$ inventory systems with server vacation under continuous and discrete time setup. Apart from deriving the explicit expressions for several measures of system performance, we have shown numerically the optimum values for $s$ and $S$. More specifically, this thesis is a study of stochastic models of both continuous and discrete time $(s, S)$ inventory system with single vacation, multiple vacation, modified multiple vacation and modified $M$ vacation policies, in which existing assumptions in the literature
are related to making the models more realistic.

Since a server adopts a different vacation policy in each model, we investigate the best vacation policy, among other vacation policies, by numerically comparing its optimal cost function of each model. The cost functions which are derived in this work are non-linear and complex, and hence it is an arduous task to determine the optimal values of cost and parameters analytically. Hence we carried out a detailed analysis with the help of numerical examples, which provides valuable insight into the operation of the system.

This thesis contains nine chapters. **Chapter 1** is an introduction to the literature and ideas surrounding the inventory with retrial demand and postponed demands along with server vacation under discrete and continuous time setup. **Chapter 2** set up the basic notations and establishes the same results which are used throughout the thesis. It contains a preliminary consideration of discrete time Markov chain, continuous time Markov chains, phase type distribution, Markovian arrival process, Matrix analytical methods and some basic system performance measures.

**Chapter 3** considers continuous review \((s, S)\) inventory systems with retrial demands and server vacation in which demands arrive according to a Poisson demand. The lead time for the reorder is assumed to be distributed as exponential. During the stock period and or server vacation period, any arriving demands enter into an orbit of infinite size. Demands in the orbit retry for their demands after a random time, and this inter-retrial time follows an exponential distribution. The server takes a vacation during the stock-out period. In this chapter, we formulate two models, namely, one is a retrial inventory system with single vacation policy and the another one is a retrial inventory system modified multiple vacation policy. Each vacation policy differs in the way that the server goes for vacation. We assume the vacation time follows an exponential distribution.

- **Single vacation policy:** Whenever the inventory level becomes zero, server goes for a vacation. Upon the return from a vacation, if ordered items are not replenished, the server remains dormant in the system until replenishment occurs.

- **Modified multiple vacation policy:** Whenever the inventory level becomes zero, the server remains idle in the system for a random period of time called (idle period) and this idle period follows an exponential distribution. If ordered items are not received during the idle period, the server goes a vacation. Upon return from the vacation, replenishment doesn’t occur, again server remains idle in the
system. This idle time and vacation time alternate until replenishment occurs in the system.

The joint probability distribution of the number of demands in the orbit, server status and on hand inventory level is derived using matrix geometric methods. Some system performance measures are obtained in the steady state case. We investigate the best model among the multiple vacation policy of Sivakumar [2011] and above two models (single and modified multiple vacation policy) numerically and illustrates some numerical examples.

In **Chapter 4**, we analyse a single server inventory system with postponed demands and various vacation policies. We assume a Markovian arrival process for the arrival of demands and independent exponential distribution for lead time, vacation time, idle time and inter-selection time. During stock out period and or server vacation period, any arriving demands enter the pool of infinite size. As long as inventory level is greater than the reorder level, pooled demands are selected one by one, according to FCFS rule. We formulate three models, each model has a different vacation policy, such as modified multiple vacation, single and multiple vacation in a postponed inventory system and compare these three models numerically. Limiting probability distributions of each model are obtained in the steady state. The total expected cost function is calculated by using system performance measures. We justify the best vacation policy, by comparing the optimal total expected cost rate of each model numerically.

**Chapter 5** is a comparative study of a Markovian inventory system with postponed demands and various vacation policies, which is extended work of chapter 4, by assuming a non-identical phase-type distribution for lead time, vacation time and inter-selection time. We calculate the expected total cost rate of each model using its system performance measures. We compare these three vacation policies numerically and present some other numerical examples.

In **Chapter 6**, we study the above vacation policies to a discrete time inventory system with postponed demands. The arrival process follows a discrete Markovian arrival process. The ordering policy consider in this chapter is \((s, S)\) policy with geometrically distributed lead time. The idle time, the vacation time and the inter-selection time are assumed to follow independent and non-identically discrete phase type distribution. The joint probability distribution of the number of demands in the pool, the server status and the on hand inventory level is obtained in the steady state case. Some system
performance measures in the steady state are derived to obtain the total expected cost function. We compared and justified the best vacation policy numerically and presented some illustrated numerical examples.

Chapter 7 introduces a modified $M$ vacation policy in a postponed inventory system with infinite pool size. Here we combined both modified multiple vacation policy, which is considered in chapter 3, and the $J$ vacation policy, which was considered by Ke [2005], as modified $M$ vacation policy. We assume a single server $(s, S)$ inventory system in which server may take at most $M$ inactive period, which comprises inactive idle period and vacation period whenever the inventory level is empty. After completing the $M$ inactive period, when the inventory is still zero, the server remains dormant in the system until the replenishment occurs. The demands occur according to Markovian arrival process. We assume independent phase type distributions for the inactive idle period and the vacation period of the server. We also assume exponential distribution for the inter selection time. We derive some system performance measures on a steady state case and some illustrative examples are discussed.

Chapter 8 deals with a modified $M$ vacation policy in a postponed inventory system with finite source. We assume a finite number of $N$ identical sources generates the arrival process. We assume $(s, S)$ ordering policy with phase-type lead time distribution. We assume phase-type distribution for the inactive idle time and the vacation time. We also assume the inter-selection time follows an exponential distribution. The joint distribution of the mode of the server, server status, the inventory level and the number of demands in the pool is obtained in the steady state case. We derived several system performance measures and total expected cost function under a suitable cost structure. The sensitivity analysis of the cost and parameter on the total expected cost function is obtained.

We present the summary of the thesis in the final chapter.