KASDI MERBAH UNIVERSITY
OUARGLA

Faculty of Mathematics and Material
Sciences

N° d'ordre :
N° de série :

Department of Mathematics

DOCTORAL DISSERTATION

Specialty: Mathematics

Option: Probabilies and Statistics

Present by: Ines Ziad

Theme

ANALYSIS OF MARKOVIAN SYSTEMS WITH WORKING VACATIONS

Represented in: 18/04/2024

Jury members:

Chairman:	Mabrouk MEFLAH	Prof.	UKM , Ouargla
Supervisor:	Amina Angelika BOUCHENTOUF	Prof.	UDL, Sidi Bel Abbes
Co-Supervisor:	Shakir MAJID	A. Prof.	Univ. Lovely Professional, India
Examiner:	Lahcen YAHIAOUI	A. Prof.	UDMT, Saida
Examiner:	Mohammed CHIKHI	Prof.	UKM , Ouargla
Examiner:	Abdelrazzak GHAZAL	Prof.	UKM , Ouargla

University year: 2023/2024

Dedication

This thesis is dedicated to my family, who have been my constant source of support and encouragement. I am especially grateful to my parents, who have inspired me in my education. I also thank my husband, Mohammed Nader Maamir, for his love and support throughout this journey. Finally, I express my love and appreciation to my daughter, Laurine, who brings joy to my life.

Acknowledgments

First and foremost, I praise and thank "ALLAH", the Almighty and Most Merciful, for granting me the patience and perseverance to complete this work.

I would like to express my sincere gratitude to my thesis supervisor, Professor Amina Angelika Bouchentouf, for her guidance, continuous support, and patience throughout my PhD studies. I have learned a lot from her and I appreciate her invaluable feedback. I also thank my co-supervisor Doctor Shakir Majid for his assistance and advice.

I am grateful to the members of the committee who agreed to examine my thesis and provide me with constructive comments and suggestions.

I would also like to acknowledge the help and support of everyone who has contributed to this work, locally or internationally. I especially thank my husband, Mohammed Nader Maamir, for his love and encouragement during the past five years of study. He has been my pillar of strength and motivation.

ملخص

تسعى الرسالة إلى إدخال نماذج انتظار جديدة لمعالجة تحديات عمليات مراكز الاتصال. تم تحليل نموذجي انتظار ماركوفيين بنجاح مع تبني سياسة العمل خلال الإجازات، مما يسمح للخوادم بأداء مهام ثانوية بمعدلات خدمة منخفضة عندما يكون النظام فارغًا. يعتبر هذا النهج مهمًا لتحسين أداء مراكز الاتصال وضمان تجربة عملاء سلسة وفعالة.

الأهداف الرئيسية لهذا البحث تشمل استخراج احتمالات الحالة الثابتة ومقاييس الأداء لهذه النظم باستخدام الأسلوب المصفوفي الهندسي. بالإضافة إلى ذلك، نقوم بتطوير نماذج التكلفة وإجراء تحليلات حساسية لتقييم تأثير مختلف معلمات النظام على مقاييس الأداء بشكل شامل.

النموذج الأول الذي تم فحصه هو نظام انتظار $M/M/c$ الذي يشمل خوادم انتظار، ورفض، وتراجع، وإجازات نشطة متغيرة K ، ويخضع لانقطاع إجازة جدول برنولي. بعد حساب الاحتمالات الثابتة للنظام، تم تطوير نموذج التكلفة باستخدام طريقة البحث المباشر لتحديد معدلات الخدمة المثلى خلال فترات العمل والإجازات بهدف تقليل التكاليف.

النموذج الثاني يدرس نظام انتظار $M/M/c$ مع إجازات نشطة متباينة (DWV) ، واستخدام جدول برنولي لانقطاع الإجازات النشطة ونفاد صبر العملاء. يركز النموذج على إجازات طويلة (نوع ١) بعد خدمة عميل وإجازات قصيرة (نوع ٢) بعد العودة. بعد الحالة المستقرة، يتم تحسين الأداء بواسطة دالة التكلفة وتحليل للتكاليف والإيرادات والأرباح، لضمان التوازن بين جودة الخدمة وكفاءة النظام وتكاليف التشغيل.

الكلمات المفتاحية: نماذج الطابور الماركوفية، إجازات نشطة، مقاييس الأداء، نفاد صبر العملاء، نموذج التكلفة.

Abstract

The main contribution of this thesis is the introduction of novel Markovian queueing models that capture the complex dynamics of call center operations. These models use the working vacation policy, which allows servers to perform secondary tasks at reduced service rates when the system is idle, a common and effective strategy in call centers, where servers can handle different types of calls or other activities during idle periods.
The main objectives of this research encompass deriving steady-state probabilities and performance measures for these queueing systems, utilizing the matrix geometric method. Additionally, we develop cost models and conduct sensitivity analyses to comprehensively assess the impact of various system parameters on performance metrics.
At first, we examine an $M/M/c$ queue with waiting servers, balking, reneging, and K-variant working vacations, subject to Bernoulli schedule vacation interruption. The steady-state probabilities of the system are obtained, and the cost function is optimized by a direct search method to find the optimal service rates during during working vacation and regular working periods.
Then, we investigates an $M/M/c$ queueing system with differentiated working vacations (DWV), Bernoulli schedule working vacation interruption, waiting servers, and customer impatience (balking and reneging). This model introduces the concept of differentiated vacations, consisting of long vacations, taken after serving at least one customer (type-1 vacation), and short vacations, taken immediately after returning from the previous vacation with an empty system (type-2 vacation). We obtain the steady-state solution, formulate a cost function and perform cost-revenue-profit analysis.

Keywords: Markovian queueing models, working vacation, performance measures, customer impatience, cost model.

Résumé

La principale contribution de cette thèse est l'introduction de modèles de files d'attente markoviens innovants qui capturent la dynamique complexe des opérations de centres d'appels. Ces modèles utilisent la politique de vacances actives (working vacation), qui permet aux serveurs de faire des tâches secondaires à des taux de service réduits quand le système est inactif, une stratégie courante et efficace dans les centres d'appels, où les serveurs peuvent gérer différents types d'appels ou d'autres activités pendant les périodes d'inactivité.

Les objectifs principaux de cette recherche sont de dériver les probabilités à l'état stable et les mesures de performance pour ces systèmes d'attente, en utilisant la méthode de matrice géométrique. Nous développons aussi des modèles de coûts et faisons des analyses de sensibilité pour évaluer l'impact de divers paramètres du système sur les mesures de performance.

En premier lieu, nous considérons un système de files d'attente $M/M/c$ avec vacances actives (working vacation), interruption de vacances selon une loi de Bernoulli, et clients impatients (balking et reneging). Ce système permet aux serveurs de prendre des vacances chaque fois que le système est vide après une période d'attente aléatoire (waiting servers), avec un nombre maximal de K vacances consécutives. Les probabilités à l'état stable du système sont obtenues, et la fonction de coût est optimisée par une méthode de recherche directe pour trouver les taux de service optimaux pendant les périodes de vacances actives et de travail régulier.

En second lieu, nous étudions un système de files d'attente M/M/c avec vacances actives différenciées, interruptions de vacances actives selon une loi de Bernoulli, waiting servers, et clients impatients (balking et reneging). Ce modèle introduit le concept de vacances actives différenciées, comprenant des vacances longues, prises après avoir servi au moins un client (vacances actives de type 1), et des vacances courtes, prises immédiatement après le retour des vacances précédentes avec un système vide (vacances actives de type 2). Nous obtenons la solution à l'état stable, formulons une fonction de coût et effectuant une analyse coût-revenu-profit.

Mots clés: Modèles de files d'attente Markoviens, vacances actives, mesures de performance, clients impatients, modèle de coûts.

List of works

Numerous research works were conducted as part of the development of this PhD thesis.

List of research works

1. Ziad, I., Laxmi, P. V., Bhavani, E. G., Bouchentouf, A. A., Majid, S. (2023). A Matrix geometric solution of a multi-server queue with waiting servers and customers' impatience under variant working vacation and vacation interruption, *Yugoslav Journal of Operations Research*, 33(3), 389-407.

2. Bouchentouf, A. A., Viajya Laxmi, P., Ziad, I. (2023). $M/M/c/DWV$ queueing model with Bernoulli schedule working vacation interruption, waiting servers, and impatience as a model of a call center, *submitted.*

Presentations

1. Ziad, I., Bouchentouf, A. A., Guendouzi, A. (May 26 - 27, 2021). $M^X/M/1$ queueing system with waiting server, K-variant vacations and impatient cutomers, *Poster presentation at the 1st International Conference on Pure and Applied Mathematics*, organized by Kasdi Merbah University of Ouargla, Algeria.

2. Ziad, I., Bouchentouf, A. A., Guendouzi, A. (November 08 - 09, 2022). On $M/M/1$ queue with differentiated vacation and impatient customers, *Poster presentation at the National Conference New Trends in Theoretical and Computational mathematics*, organized by Amine Okkal Hadj Moussa Eg Akhamouk University of Tamanghasset, Algeria.

Contents

List of works **v**

Introduction **1**

1 Queueing theory: Historical background, basic concepts, Markovian queues, and impatience **4**
1.1 A brief historical facts on queueing theory 4
1.2 Applied examples of queueing theory 5
1.3 Basic concepts and terminology of a queueing system 6
1.4 Kendall notation for classifying queues 7
1.5 Performance measures in queueing systems 8
1.6 An introduction to Markovian queueing models 9
1.6.1 Markovian single-server models 9
1.6.2 Multi-server Markovian models 10
1.7 Customer impatience in queueing systems 11

2 On working vacation queueing models **12**
2.1 Types of vacation policies in queueing theory 12
2.2 Working vacation policy 13
2.3 Applications of WV policy 14
2.4 A review of some working vacation models in queueing theory 15
2.4.1 An $M/M/1$ and $M/G/1$ queueing models with WV and various extensions 15
2.4.2 Generalized input models with WV 16
2.4.3 WV models with Markov Arrival Processes 17

3 On matrix analytic method **18**
3.1 Introduction 18
3.2 Diverse applications of matrix analytic methods 19
3.3 Utilizing matrix analytic methods 20
3.3.1 Building the generator 20
3.4 Quasi Birth-and-Death processes (QBD) 25
3.4.1 Matrix-geometric method 26

4 A matrix geometric solution of a multi-server queue with waiting servers and customers' impatience under variant working vacation and vacation interruption **29**
4.1 Introduction . . . 29
4.2 The model . . . 31
4.2.1 Practical application . . . 32
4.3 Analysis of the model . . . 33
4.3.1 Matrix-geometric method . . . 33
4.4 Performance measures and cost model . . . 38
4.4.1 Performance measures . . . 38
4.4.2 Cost model . . . 39
4.5 Some special case . . . 40
4.6 Numerical illustrations . . . 40
4.7 Conclusion . . . 45

5 $M/M/c/DWV$ queueing model with Bernoulli schedule working vacation interruption, waiting servers, and impatience as a model of a call center **46**
5.1 Introduction . . . 46
5.1.1 Motivation of the current research work . . . 47
5.2 Model description and practical application . . . 48
5.2.1 Model description . . . 48
5.2.2 Practical justification of the proposed queueing model 49
5.3 Analysis of the steady-state probability distribution . . . 50
5.3.1 Infinitesimal generator of the process . . . 53
5.3.2 Equilibrium condition of the system . . . 54
5.3.3 Stationary probability distribution . . . 55
5.4 System performance measures . . . 56
5.5 Special cases of our model . . . 57
5.6 Numerical analysis . . . 58
5.6.1 Cost-revenue-profit modeling . . . 58
5.6.2 Parameter impact study . . . 59
5.6.3 Discussion . . . 59
5.7 Conclusion . . . 65

Conclusion and future directions **65**

Bibliography **69**

Introduction

Many real-world systems involve waiting lines of customers or requests that require service from one or more servers. Examples of such systems include communication networks, computer systems, manufacturing systems, service systems, etc. Queueing theory is a branch of applied mathematics that studies the behavior and performance of these systems, and provides insights into their operational characteristics, such as the average number of customers in the system, the average waiting time of customers, the probability of blocking or loss, the utilization of resources, etc. (see [56, 179]).

One of the most important and widely used classes of queueing models is the Markovian queueing models, which assume that the interarrival times and the service times of customers follow exponential distributions. Markovian queueing models have the advantage of being mathematically tractable and analytically solvable, and can capture many essential features of real queueing systems. Markovian queueing models can be further classified into different types, depending on the number of servers, the number of phases, the buffer size, the service discipline, the feedback mechanism, etc.

In this thesis, we aim to analyze and optimize Markovian queueing systems with working vacations and impatience. Working vacations are periods when the servers can perform some secondary tasks when the system is empty, and resume normal service when the system is nonempty. This can be useful for modeling systems where the server has some secondary tasks to perform when there are no customers in the queue, such as maintenance, backup, or updating. This policy can also reduce the waiting time and the queue length of the customers, and improve the system performance and efficiency (see [22, 36]). Impatience refers to the phenomenon that customers may leave the system before being served due to their limited waiting time or tolerance. Impatience can affect the customer satisfaction and loyalty, as well as the revenue and profit of the service provider.

To address this goal, we analyze different working vacation queueing models based on the matrix-analytic method (MAM), a powerful technique for deriving analytical solutions for complex queueing systems, obtaining the steady-state probabilities and performance measures of the systems. This method uses matrix operations and linear algebra to solve various queueing models. For recent advances on the theme see [126, 37, 41].

The thesis is structured as follows:

Chapter 1 introduces the basic concepts of queueing theory, with a focus on Markovian queues and impatience behaviour in queueing systems.

Chapter 2 reviews the literature on working vacation queues, which are queueing systems where the servers can perform some secondary tasks when the system is empty.

Chapter 3 presents the matrix-analytic method, a powerful technique for solving various queueing models that are difficult to analyze by traditional methods.

Chapter 4 analyzes a $M/M/c$ queueing system with waiting servers, balking, reneging, and K-variant working vacations subjected to Bernoulli schedule vacation interruption. This model captures a realistic scenario where the servers can take multiple types of vacations with different service rates depending on the system state. The customers can decide to join or leave the queue based on the queue length and their patience level. The vacations can also be interrupted by the arrival of customers with a certain probability. K-variant vacation is a generalization of single and multiple vacation policies, and it has been studied by Vijaya Laxmi and Rajesh [105], Vijaya Laxmi et al. [170], Bouchentouf and Guendouzi [30]. This model can be applied to call centers, where the agents can handle different types of calls, and perform some secondary tasks when there are no calls in the queue.
We use the matrix-geometric method to obtain the steady-state probabilities of the system, and derive important performance measures of the queueing model, such as the mean number of customers, the average number of loss customers due to balking and reneging, etc. We also construct a cost model and apply a direct search method to find the optimum service rates during both working vacation and regular working periods at lowest cost. We provide numerical results to illustrate the effects of various parameters on the system performance and the optimal solution. This chapter is based on a paper that has been published in Yugoslav Journal of Operations Research, 33(3), 389-407, 2023.

Chapter 5 investigates a $M/M/c/DWV$ queueing model with Bernoulli schedule working vacation interruption, waiting servers, and impatience as a model of a call center. This model considers two types of vacations: long and short vacations. The long vacation is taken after serving at least one customer (type-1 vacation), while the short vacation is taken immediately after returning from the previous vacation and finding the system empty (type-2 vacation). This differentiated vacation (DV) approach has been recently introduced by Ibe and Isijola [71] and Isijola and Ibe [72], and it has many applications in real-life situations such as call centers, power save/sleep mode for energy-efficient utilization in modern mobile technologies, hospital emergency room operation, gas stations, etc. Readers interested in this topic can refer to Bouchentouf et al. [33], Suranga Sampath and Liu [152].

For the proposed queueing model, we build a Markovian process of the considered queueing system. Then, by employing the matrix geometric method, we obtain the steady-state probabilities. We also derive various performance measures in a steady state. In addition, we carry out the sensitivity analysis to illustrate the impact of different system parameters on the performance characteristics, which can provide insight to the system managers to supervise the operation status of this system and reduce the congestion problem. Moreover, we develop a cost function to help the system managers or decision-makers regulate the system economically. This chapter is based on a paper that has been submitted.

The concluding section of the thesis summarizes the main contributions and results presented, and discusses their implications and limitations. It also provides some directions for future works.

Chapter 1

Queueing theory: Historical background, basic concepts, Markovian queues, and impatience

In this chapter, we present a brief overview of queueing theory. At first we present some historical facts. Then, we introduce the basic concepts and notation of a queueing system, we present a brief introduction on Markovian models, where we present some examples. We then explore the phenomenon of customer impatience, as this concept has been well studied in this thesis. We focus on Markovian queues and customer impatience because they are the most common and widely used models in queueing theory, and they have many interesting and challenging properties and applications queueing is a common phenomenon in our daily life.

1.1 A brief historical facts on queueing theory

The origin of queueing theory dates back to 1917, when the Danish engineer and mathematician Agner Krarup Erlang investigated the management of telephone networks in Copenhagen. He developed models for queue access systems, different priorities for arrivals, and statistical distributions of service times. Later, many mathematicians contributed to the advancement of the theory, such as Khinchin, Palm, Kendall, Polachek and Kolmogorov. Kendall established the foundations for the analysis of queueing systems using compact Markov chains and introduced a notation for queueing systems that is still used today. Lindley derived a formula that allows for the calculation of the performance of a queue system under general conditions. Not long after (in 1957), Jackson developed the so-called queueing networks. With

the development of computers and computer networks, queuing systems and queueing networks became a powerful tool for analyzing and designing various applications.
The mathematical basis for the analysis of queueing networks was provided by Whittle [177, 178] and Kingman [88], who used the terminology of population processes. Complex queueing network problems have been extensively researched since the early 1970s. Several surveys and books summarize the main results in this area, such as Guarguaglini et al. [57], Kelly [85], Whitt [175, 176], Harrrison and Nguyen [65, 66] and Dai [42]. Bhat and Basawa [21] used both queue length and waiting time data to estimate parameters in queuing systems. Other researchers used the matrix analytic method, such as Alfa [3], and Lothar and Dieter [119].
During the last century, many important books on queueing theory and its applications have been published, such as those by Kleinrock [89, 90, 91], Daigle [43], Anisimov [7], Haghighi [60], Shortle et al. [146], and Yue et al. [192], to name a few.

1.2 Applied examples of queueing theory

Queueing theory is a useful tool for analyzing and optimizing various situations where queues occur. Some examples of applications and principles are:

- Banking: Banks are one of the main places where people encounter queues. Queueing theory can help determine the optimal number of servers to minimize the waiting time and cost. It can also help customers to choose the best service option and arrival time based on the queue length and service rate.

- Hospitals: Hospitals often face long queues due to high demand and limited capacity. Queueing theory can provide important measures of performance for servers and queues, such as the average waiting time, queue length, and utilization. It can also suggest ways to improve the service efficiency and quality, such as by increasing the service rate, prioritizing patients, or allocating resources. Moreover, queueing theory can model complex situations where patients move across different departments or services.

- Computer systems and communication networks: Queueing theory has been widely used to study the performance of computer systems and communication networks. Simple or analytical queueing models can provide a cost-effective way to estimate the reliability and performance metrics of these systems, such as the throughput, response time, and availability. For example, queueing models can help identify and eliminate the bottleneck components in a computer system, such as the CPU, memory, disk, or network interface.

- Transportation: Queueing theory can also be applied to transportation systems, such as roads, railways, airports, or ports. Queueing models can help to analyze the traffic flow, congestion, delay, and safety of these systems. They can also help to design and manage the transportation infrastructure, such as by optimizing the traffic signals, routing strategies, toll policies, or parking facilities.

- Manufacturing: Queueing theory can be used to model and optimize manufacturing systems, such as production lines, assembly lines, or supply chains. Queueing models can help to measure and improve the productivity, quality, and profitability of these systems. They can also help to plan and control the production process, such as by scheduling the jobs, allocating the machines, or managing the inventory.

1.3 Basic concepts and terminology of a queueing system

Queueing systems are mathematical models that describe the behavior of systems where customers arrive, wait, and receive service. Queueing systems can be characterized by several factors, such as:

- The arrival process: This defines how customers arrive and enter the system. It is usually described by the inter-arrival time between consecutive arrivals or the arrival rate per unit of time. The inter-arrival times are often assumed to be random and independent, following a certain probability distribution. The customers may come from an infinite or a finite source, and they may arrive individually or in batches of variable or fixed size.

- The service process: This defines how customers are served by the system. It is usually described by the service time for each customer or the service rate per unit of time. The service times are often assumed to be random and independent, following a certain probability distribution. The service process may depend on the state of the system, such as the number of customers waiting or being served.

- The queue discipline: This defines the rule that determines which customer is selected for service from the waiting line. The most common queue discipline is First-Come First-Served (FCFS) or First-In First-Out (FIFO), but there are other possibilities, such as last-in first-out (LIFO), random order (SIRO), priority order, or service in groups.

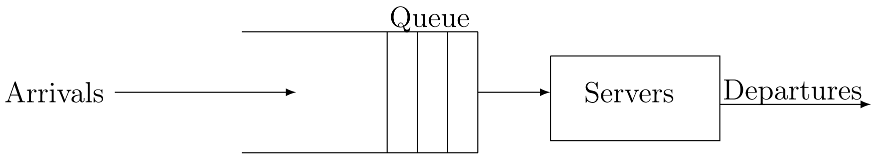

Figure 1.1: Basic queueing model.

- The capacity of the system: This defines the maximum number of customers that can be in the system at any time. Some systems may have a finite capacity, which means that some customers may be rejected or lost if the system is full. Other systems may have an infinite capacity, which means that no customer is ever turned away.
- The number of servers: This defines the number of parallel channels that can serve customers simultaneously. A queueing system is called a single-server model when it has only one server, and a multi-server model when it has more than one server.

These factors affect the performance and efficiency of queueing systems, which can be measured by various indicators, such as the average waiting time, queue length, utilization, or probability of loss.

1.4 Kendall notation for classifying queues

Kendall notation is a standard way of classifying queueing models using symbols that describe their main characteristics. Kendall introduced this notation in 1953, as follows:

arrival process/service distribution/number of servers/buffer size-queue discipline.

The symbols used for the first two positions are usually: D (deterministic), M (Markovian or Poisson for arrivals or exponential for service times), G (general), GI (general and independent), and $Geom$ (geometric). The third position indicates the number of servers in the system. The fourth position indicates the capacity of the system, including the servers and the waiting space. If the system has infinite capacity, this position is omitted. The fifth position indicates the queue discipline, which is the rule that determines the order of service for customers. The most common queue disciplines are: FIFO (first-in first-out), LIFO (last-in first-out), PS (processor sharing), and Random. If the queue discipline is FIFO, this position is omitted. A dash sign "-" is used before the fifth position to avoid confusion if the fourth position is missing.

For example, $M/M/1$ means a single-server queue with Poisson arrivals, exponential service times, infinite capacity, and FIFO order. $M/G/K/K$ means a K-server queue with Poisson arrivals, general service times, no waiting space, and FIFO order. $M/G/1 - PS$ means a single-server queue with Poisson arrivals, general service times, infinite capacity, and processor sharing order. Note that in an $M/G/1-PS$ queue, the service time of a customer starts immediately upon arrival, but it may last longer than its service requirement, because the server capacity is shared among all customers in the system.

1.5 Performance measures in queueing systems

As we navigate the world of queue analysis, it becomes clear that we need to assess how well these systems perform. For that purpose, several key performance measures are recommended:

- Average customer queue and system size: This metric focuses on the average number of customers waiting in both the queue and the entire system.
- Average waiting time: It quantifies the average time customers spend waiting in the queue and throughout the system.
- System utilization rate: This reflects the percentage of server capacity in active use.
- Cost of service level: An evaluation of the expenses associated with the implemented service capacity.
- Probability of customer wait: The likelihood that a prospective customer will experience wait times before being served.
- Traffic intensity: A comprehensive measure that encompasses various aspects of the system's behavior.

Among these, the system utilization rate merits particular attention. Instead of merely indicating server inactivity, it reveals the extent to which servers are occupied. The concept of achieving 100% utilization, intuitively appealing for efficient resource management, has its nuances. As the utilization rate approaches 100%, both the number of waiting customers and the average waiting time tend to increase indefinitely. Full server occupancy leads to customer waits, contrary to the 100% utilization ideal, especially during peak business hours. This highlights the imperative for achieving a balanced system where the combined cost of service and waiting remains minimal.

1.6 An introduction to Markovian queueing models

Markovian queueing models are a particular type of system where inter-arrival times and service times are random variables following exponential distributions, and they are independent of each other. These models are particularly amenable to study due to the memoryless property of the exponential distribution.

1.6.1 Markovian single-server models

Let's explore two common types of Markovian single-server queueing systems:

- $M/M/1/\infty$ **queueing system:**
  1. The first "M" represents Markovian or memoryless arrivals.
  2. The second "M" denotes memoryless service times, following exponential distributions.
  3. "1" signifies a single server in the system.
  4. "∞" implies that the system has infinite capacity, meaning it can accommodate any number of customers. Figure 1.2 illustrates this model.

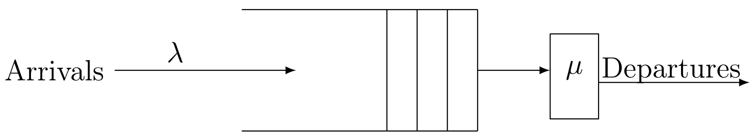

Figure 1.2: The $M/M/1$ queueing model.

The transition diagram of the $M/M/1$ queue is depicted in Figure 1.3.

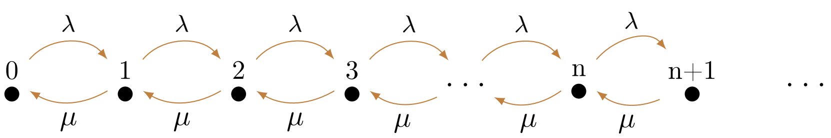

Figure 1.3: Transition diagram of an $M/M/1$ queuing system.

- $M/M/1/N$ **queuing system:**
  This is an extension of the $M/M/1/\infty$ queue where the system can have at most "N" customers in the queue.

1.6.2 Multi-server Markovian models

In Markovian queueing theory, we also encounter models with multiple servers. Let's present a few of these:

- $M/M/c/\infty$ **queueing system**: In this Markovian queueing model, the system features "c" servers to handle customer requests (see Figure 1.4).

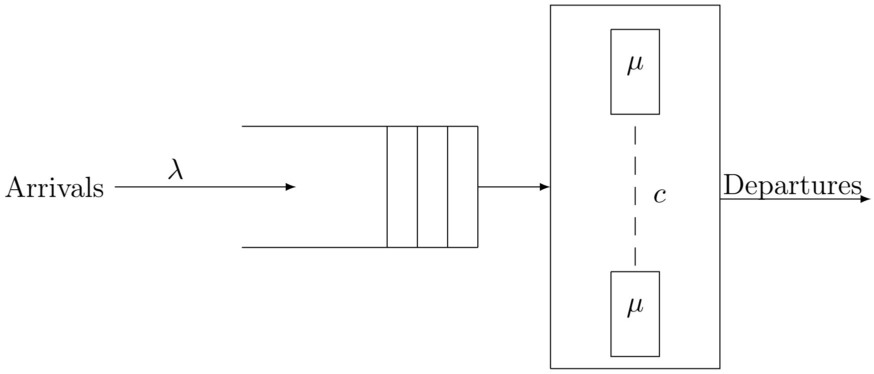

Figure 1.4: The $M/M/c$ queueing model.

The transition diagram of the $M/M/c$ queue is presented in Figure 1.5.

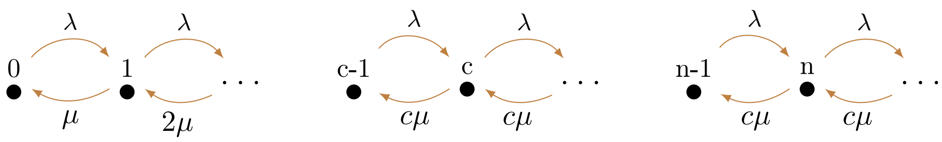

Figure 1.5: Transition diagram of an $M/M/c$ queuing system.

- $M/M/c/N$ **queueing system:**
  - This model introduces a system with finite capacity, denoted as "N."
  - It's important to note that in this scenario, "c" (the number of servers) is less than "N" (the system's capacity).

- $M/M/\infty$ **Queuing System:**
  - This Markovian queueing system is unique because it operates without a queue.

- An infinite number of servers are available, ensuring that every arriving customer immediately finds an idle server.

1.7 Customer impatience in queueing systems

Customer impatience is a realistic phenomenon that affects the performance of queueing systems, as customers may abandon the queue before being served due to frustration or dissatisfaction. Different types of impatience behaviors have been considered in the literature, such as reneging, balking, jockeying, and retrial:

- **Reneging**: This occurs when a customer who has joined the queue decides to leave it before getting service. Typically, the customer makes this decision based on their perceived waiting time exceeding their tolerance level.

- **Balking**: This refers to the situation where a customer chooses whether to join the queue or not depending on their expected waiting time. The customer may also make this decision after observing the queue length or being informed of the waiting time.

- **Retrial**: In addition to balking and reneging, there is another impatient behavior called retrial. This is a phenomenon where a customer who may balk at entering the system or renege from it may join a virtual queue of blocked customers and retry for service after a random time interval. A queueing system with retrial phenomenon is called a retrial queue. The probability that a customer may balk or renege and enter the orbit depends on the number of customers being served. Retrial queues have applications in various domains, such as e-commerce, computer systems, communication networks, etc.

- **Jockeying**: This involves customers switching between queues in a multi-server system in order to minimize their waiting time.

Haight [64] was the first to introduce the idea of customer impatience in queueing theory. He studied a $M/M/1$ queue with balking behavior. He also examined a queue with reneging behavior. Since his pioneering work, there has been a lot of research on queueing systems with balking and reneging. Some of the recent works on this topic are Wang et al. [173], Danilyuk et al. [45], Sudesh and Azhagappan [150], and Keerthiga and Indhira [84].

Chapter 2

On working vacation queueing models

In the field of queueing theory, there has been extensive research on queues with server vacations, which has been a topic of interest for a long time. Server vacation models have gained significant attention since Levy and Yechiali's work in 1975 [108]. In these models, the server may become temporarily unavailable for serving customers. This unavailability, often utilized for different purposes like maintenance, attending to other queues, or taking breaks, is referred to as a "vacation."

Queueing systems with server vacations have found widespread applications in various domains such as computer communication systems, manufacturing and production systems, and inventory management. For example, in computer networks, server vacations can be used to model the behavior of routers that switch between different transmission modes or channels. In manufacturing systems, server vacations can be used to model the behavior of machines that undergo preventive maintenance or breakdowns. In inventory management, server vacations can be used to model the behavior of retailers that replenish their stock from a supplier.

For more in-depth insights into this topic, monographs, books, and comprehensive surveys on server vacation models can be found in the queueing literature, such as those by Doshi [48], Takagi [154], and Tian and Zhang [159], Tian et al. [158] and Ke et al. [82] along with the references they provide.

2.1 Types of vacation policies in queueing theory

A vacation policy is a rule that determines when and how the server in a queueing system takes a break from serving customers. Studying different vacation policies can help to understand and optimize the performance and efficiency of queueing systems under various scenarios and constraints. In queueing theory, various vacation policies have been studied, such as:

1. **Single vacation queueing models**: The server goes on vacation at the end of each busy period and returns immediately after the vacation period, regardless of the system's status.

2. **Multiple vacation queueing models**: The server goes on a series of vacations until he finds that the system is not empty when he completes a vacation period.

3. **Working vacation queues**: The server works continuously during this period but at a reduced rate.

4. **Gated vacation queues**: The server serves only customers within a "gate" placed behind the last waiting customer, based on certain rules.

5. **Limited service discipline**: The server proceeds on vacation after serving K consecutive customers, after a predefined time interval, or when he becomes idle.

6. **Exhaustive service discipline**: The server serves customers until the system is empty, after which they take a vacation of random length.

7. **Differentiated vacation queues**: In this scenario, the server has the option to take two types of vacations. The first type of vacation occurs after serving at least one customer, while the second type is taken when the server returns from the previous vacation and finds that the system is empty.

8. **K-variant vacation queues**: At the moment when a vacation is completed, if there are customers waiting in the queue, the server switches to a busy period. However, if there are no customers in the queue, the server has the option to take up to K consecutive vacations before eventually returning to the busy period. During the busy period, the server remains active until new customers arrive.

2.2 Working vacation policy

Working vacation (WV) is a vacation policy in which the server continues to provide service during the vacation period, albeit at a reduced rate, instead of halting service entirely. The concept was first introduced by Servi and Finn [145]. Working vacation can improve the system performance by reducing customer waiting times, queue lengths, and server idle times. However, it is worth pointing out that this policy also introduces some challenges such as determining the optimal vacation rate, balancing the trade-off between service quality and energy consumption, and coordinating multiple servers with different vacation policies.

In working vacation models, the server cannot return to the regular busy period until the vacation period concludes. Nevertheless, during the WV period, if there are customers present at a service completion instant, the server has the option to interrupt the vacation and return to the regular busy state. This policy is referred to as vacation interruption. Vacation interruption can further improve the system performance by reducing customer waiting times and queue lengths. However, it also introduces some challenges such as determining the optimal interruption probability, balancing the trade-off between service quality and server fatigue, and coordinating multiple servers with different interruption policies.

2.3 Applications of WV policy

The concept of WV policy is particularly relevant for modeling systems where the server has secondary tasks to perform when the queue is empty, such as maintenance, backup, or updates. The models incorporating WV have found applications in various fields, including some examples given bellow:

- Call Center with multi-task employees: Employees can handle various call types, each with different priorities and skill requirements. During idle times, they can engage in secondary tasks like training, coaching, or administrative work and switch between normal service and working vacation as per predefined rules.

- Customized manufacturing system: This system produces customized products based on customer orders. During idle periods, the system can perform preventive maintenance or quality improvement tasks, transitioning between regular production and working vacation based on established rules.

- Telecommunication network with repairable faults: In this network, nodes can transmit data packets, and when a node fails, a repairman can repair it or engage in other tasks during fault-free periods. The repairman switches between normal repair and working vacation according to predefined rules.

- Hospital emergency department: Doctors handle different types of patients with varying priorities and service requirements. During idle times, they can engage in tasks such as checking records, writing reports, or consulting with colleagues, switching between normal service and working vacation based on predefined rules.

- Wireless communication network: Base stations transmit data packets to mobile users. During idle times, they can perform energy-saving activities like switching to a low-power mode, adjusting transmission power, or scanning the channel. They transition between normal transmission and working vacation based on predefined rules.

- Cloud computing system: Servers process various tasks with different resource demands and deadlines. During idle periods, servers can engage in load-balancing activities such as task migration, resource reallocation, or software updates, switching between normal processing and working vacation based on predefined rules.

- Air traffic control system: Controllers handle different types of flights with varying priorities and service requirements. During idle times, they can engage in tasks such as checking weather conditions, updating flight schedules, or communicating with other controllers, switching between normal service and working vacation based on predefined rules.

- Online shopping system: Servers process various orders with different payment methods and delivery options. During idle periods, servers can engage in tasks such as verifying customer information, updating inventory records, or sending confirmation emails, switching between normal processing and working vacation based on predefined rules.

- Electric vehicle charging station: Chargers provide electricity to electric vehicles with different battery capacities and charging demands. During idle times, chargers can perform energy management activities such as adjusting charging power, switching to renewable energy sources, or storing excess energy, switching between normal charging and working vacation based on predefined rules.

2.4 A review of some working vacation models in queueing theory

Several variations and extensions of working vacation queueing models have been proposed and analyzed in the literature. Some of the main contributions are summarized below:

2.4.1 An $M/M/1$ and $M/G/1$ queueing models with WV and various extensions

Several variations and extensions of $M/M/1$ and $M/G/1$ queueing models model have been proposed and analyzed in the literature. In addition to introducing a semi-vacation policy in an $M/M/1$ queue with multiple working vacations (Servi and Finn [145]), researchers extended this to an $M/G/1$

queue (Wu and Takagi [180]) and investigated the stochastic decomposition structures of queue length and waiting time in $M/M/1$ and $M^X/M/1$ queues with working vacations (Liu et al. [117]; Xu et al. [182]). Moreover, they considered a working vacation queue with multiple types of server breakdowns and repairs (Jain and Jain [74]). In addition, they proposed vacation interruption in an $M/M/1$ queue (Li and Tian [113]) and studied $M^X/M/1$ queues with single and multiple working vacations (Xu et al. [183]; Baba [13]). They also introduced randomized working vacations in an $M^X/G/1$ queue (Gao and Yao [51]). Later, researchers applied the idea of vacation interruption to an $M/G/1$ queue with multiple working vacations (Zhang and Hou [196]) and an $M/M/1$ queue with single working vacation (Majid and Manoharan [122]). Later, researchers conducted sensitivity analysis for the $M/M/1$ retrial queue with working vacations and vacation interruption (Ameur et al. [5]). They also discussed the impatience behavior of a finite buffer $M/M/1/N/DWV$ queue with vacation interruption (Bouchentouf et al. [33]). Recently, researchers conducted a steady-state analysis of a single-server $M/G/1$ queueing system with multiple working vacations (Seeniraj and Moganraj [142]) and also examined an $M/G/1$ feedback retrial queue with working vacations and a waiting server (Murugan and Keerthana [127]).

2.4.2 Generalized input models with WV

The $GI/M/1$ queue with working vacation is a common model for studying different service systems. Several variations and extensions of this model have been proposed and analyzed in the literature. In addition to investigating a general input model with working vacation (Baba [12]), researchers discussed a multiple-server model with working vacations (Banik et al. [18]) and introduced the concept of vacation interruption in a discrete-time model with geometric service time (Li and Tian [111]). Subsequently, they examined a finite buffer model with variant of multiple working vacations (Zhang and Hou [197]). Additionally, researchers analyzed a finite buffer model with state-dependent services and state-dependent multiple working vacations (Goswami et al. [53]). Furthermore, they studied a Bernoulli-schedule-controlled vacation model with vacation interruption (Tao and Wang [156]). Recently, researchers presented an analysis of a batch arrival model with multiple working vacations by using a $GI/M/1$ type Markov process (Zhang [193]). They also and conducted an in-depth examination of a discrete-time batch arrival model ($GI^X/Geo/1$ queue) with multiple working vacations under late and early arrival system (Barbhuiya and Gupta [19]) and dealt with a $GI^X/M/1$ queue with two-stage vacation policy using shift operator method (Liu et al. [118])

2.4.3 WV models with Markov Arrival Processes

Working vacation models with Markov Arrival Processes (MAP) have emerged as a recent development in the field. In addition to delving into the PH approximation for the $MAP/G/1$ queue with an N-policy working vacation (Zhao and Cui [198]), researchers investigated the $MAP/G/1$ queue with working vacations and vacation interruptions using supplementary variables along with the matrix-analytic method and censoring technique (Zhang and Hou [195]). They also introduced the concept of N-policy vacation in a model with MAP arrivals and phase-type services (Sreenivasan et al. [148]). Moreover, they examined a cold standby repairable system with working vacations and vacation interruptions under MAP arrivals (Liu et al. [116]). Later, researchers explored an $MAP/PH/1$ model with working vacations, working breakdowns, and a two-phase repair (Thakur et al. [157]).

Chapter 3

On matrix analytic method

3.1 Introduction

Matrix Analytic Methods (MAM) traces its origins to the mid-1970s, with its inception credited to Dr. Marcel F. Neuts, who championed this innovative approach primarily for the investigation of queueing models. Over the course of the past four decades, the theoretical foundations of MAM have undergone substantial development, paralleling its extensive practical applications, as substantiated by the seminal works of Neuts [131, 132], Latouche and Ramaswami [101], and He [67]). Matrix Analytic Methods comprise an indispensable suite of analytical tools essential for dissecting a diverse spectrum of Markov processes characterized by intricate structures, showcasing remarkable versatility.

These methods have become deeply ingrained in the fabric of the natural and applied sciences, engineering, as well as statistics. They play a pivotal role in modeling, performance evaluation, and the architectural design of complex systems such as computer systems, telecommunication networks, network protocols, manufacturing systems, supply chain management systems, and risk/insurance models, among others. In this chapter, we provide principles and versatile applications of matrix analytic methods (MAM), across various domains (see Section 3.2). Through meticulous analysis, this thesis aims to contribute significantly to the evolving field of matrix-analytic research, specifically in the context of applying MAM (refer to Section 3.3). The approximations developed in this thesis primarily revolve around the study of quasi birth-and-death processes using MAM, as detailed in Section 3.4. We introduce a powerful technique for determining the steady-state distribution of these processes - the matrix-geometric method, which is explained in Section 3.4.1.

3.2 Diverse applications of matrix analytic methods

Matrix analytic methods (MAM) have found widespread applications in various fields since their inception in the 1970s. This section explores the diverse domains of MAM are:

- **Queueing theory:** MAM initially emerged as a powerful tool within queueing theory, revolutionizing the analysis of queueing models. Over the years, it has been applied successfully to a wide array of queueing systems, including multi-server queues, vacation queueing models, queueing systems with customer impatience, retrial queues, and queues with server breakdowns. Typically, these studies involve the utilization of Markovian Arrival Processes (MAPs) for customer arrival processes, phase-type (PH) distributions to model service times, and various types of Markov chains such as (embedded) Quasi-Birth-and-Death (QBD), $M/G/1$, and $GI/M/1$ to describe system dynamics at specific epochs. Matrix-geometric or matrix-exponential techniques are employed to efficiently compute key queueing performance metrics. For detailed insights and references, see Neuts [131, 132], Lucantoni et al. [120], He [68], and Xia et al. [181].

- **Computer and communication networks:** MAM can be used to model and analyze various aspects of computer and communication networks, such as network traffic, routing protocols, congestion control, packet loss, network reliability, and security. Some of the common models used in this domain are Markov-modulated Poisson processes (MMPPs), Markovian fluid queues, Markovian feedback queues, and Markovian arrival networks. For more details, see Latouche and Ramaswami [101], and Nigel et al. [133].

- **Risk and insurance analysis:** MAM has made substantial inroads into risk and insurance analysis, particularly in its application to Markovian Mixture of Erlangian Fitting Factors (MMFFs). In a manner akin to queueing applications, MAM provides closed-form solutions for numerous classical risk models and associated quantities. Furthermore, efficient algorithms have been devised for calculating these quantities. This field represents a dynamic area of research within MAM in recent years. For comprehensive insights, see Asmussen [9] and Badescu and Landriault [15].

- **Inventory and supply chain management:** Owing to its resemblance to queueing systems, MAM is an effective tool for analyzing inventory and supply chain models. Beyond analysis, it aids in the development of algorithms to compute optimal control policies, significantly expanding its application scope. For in-depth information, see Chen and Song [39] and He [67].

- **Reliability theory:** In addition, MAM has also found applications in reliability theory, where PH-distributions are employed to model component failure times, and MAPs are used for events like shock-wave arrivals. The closure properties of PH-distributions and MAPs prove invaluable in assessing various reliability system performance metrics, including system availability, system time to failure, and system failure rates.

- **Other stochastic systems:** MAM can also be useful for analyzing other stochastic systems that have a matrix-analytic structure, such as stochastic Petri nets, stochastic automata networks, stochastic differential equations, stochastic games, and stochastic optimization problems. For further details, see Plateau and Atif [135] and Bini et al. [23].

3.3 Utilizing matrix analytic methods

Matrix analytic methods (MAM) represent an integration of our understanding of continuous-time Markov chains (CTMCs) and discrete-time Markov chains (DTMCs) with the intricate world of queueing systems. In this thesis, we exclusively focus on CTMCs. Thus, this section focuses on the modeling of systems in continuous time.

3.3.1 Building the generator

Continuous-time Markov chains (CTMCs) serve as a fundamental component in the application of matrix analytic methods (MAM) within queueing systems. In this part, we offer a succinct overview of CTMCs, primarily adapted from Ross [139], Chapter 6 and Granville's Thesis [55], Chapter 1.

3.3.1.1 Definition of continuous-time Markov chain

Definition 3.1. *A stochastic process $\{X(t), t \geq 0\}$ is classified as a continuous-time Markov chain if it satisfies the following criteria:*

1. *The state space $\mathcal{S}$ of $X(t)$ is, at most, countable (i.e., it can be finite, $\mathcal{S} = \{0, 1, ..., n\}$, or countably infinite, $\mathcal{S} = \{0, 1, ...\}$), rendering $X(t)$ a discrete random variable.*

2. *(Markov Property) For any $s, t \geq 0$ and $i, j \in \mathcal{S}$, the following holds:*

$$\mathbb{P}(X(t+s) = j | X(s) = i, X(u) = x(u), 0 \leq u < s)$$
$$= \mathbb{P}(X(t+s) = j | X(s) = i).$$

Here, $x(u) \in \mathcal{S}$ is the state (potentially varying) of the CTMC at some time in the past, as a function of the time u. This implies that the probabilistic characteristics governing the future evolution of the CTMC depend exclusively on its current state and remain independent of its historical trajectory.

A CTMC is labeled as time-homogeneous if $\mathbb{P}(X(t+s) = j|X(s) = i)$ is independent of s. Under such circumstances, we formulate the transition probability function $P_{i,j}(t)$ as:

$$\mathbb{P}_{i,j}(t) = \mathbb{P}(X(t+s) = j|X(s) = i) = \mathbb{P}(X(t) = j|X(0) = i), \ t \geq 0, \ i, j \in \mathcal{S}.$$

3.3.1.2 Sojourn times and absorbing states

The duration that a CTMC resides in a state $i \in \mathcal{S}$ before switching to a different state $j \neq i$ is subject to a random distribution, specifically, an exponential distribution. The rate parameter of this distribution may be state-dependent and is denoted as the sojourn time at state i, $T_i \sim \text{Exp}(v_i)$. Here, T_i is exponentially distributed with rate v_i, where v_i represents the total rate of all transitions leaving state i. If $v_i = 0$, state i is regarded as an absorbing state, with an infinite sojourn time.
Assuming $v_i > 0$, after completing the sojourn time T_i, the CTMC switches to state $j \neq i$ with probability $p_{i,j}$, where $\sum_{j \in \mathcal{S}} p_{i,j} = 1$. If only state transitions (without sojourn times) are observed, the observed movements of the stochastic process can be modeled by its embedded DTMC $\{X_n, n \in \mathbb{N}\}$ with a transition probability function matrix (TPFM) $P = [p_{i,j}]$. An observed restriction in these embedded DTMCs is that $p_{i,i} = 0$ for all $i \in \mathcal{S}$ as long as $v_i > 0$. In the case of $v_i = 0$, by convention, $p_{i,i} = 1$, making state i absorbing in both the CTMC and its embedded DTMC.
Additionally, define $q_{i,j} = v_i p_{i,j}$, $j \neq i$, as the probability flow or instantaneous transition rate from state i to state j. Consequently,

$$\sum_{j \in \mathcal{S}, j \neq i} q_{i,j} = v_i, \ i \in \mathcal{S}.$$

These definitions can be understood as follows: Let the initial distribution of a CTMC be $\alpha_0 = (\alpha_{0,0}, \alpha_{0,1}, \ldots)$, where $\alpha_{t,i} = \mathbb{P}(X(t) = i)$, $i \in \mathcal{S}$, $t \geq 0$. If $h > 0$ is a small time increment such that

$$\mathbb{P}(\geq 2 \text{ transitions in } [0, h] | X(0) = i) = o(h), \quad \forall i \in \mathcal{S},$$

where $o(h)$ denotes a function with

$$\lim_{h \to 0} \frac{o(h)}{h} = 0,$$

then, employing Taylor series expansion, we find that

$$\mathbb{P}(1 \text{ transition in } [0, h] | X(0) = i) = v_i h + o(h),$$

and for $j \neq i$,

$$P_{i,j}(h) = q_{i,j} h + o(h). \tag{3.1}$$

Similarly, we can establish that

$$P_{i,i}(h) = P(T_i > h) + o(h) = e^{-v_i h} + o(h) = 1 - v_i h + o(h). \tag{3.2}$$

Consequently,

$$\mathbb{P}(X(h) = j) = \alpha_{0,j} - \alpha_{0,j} v_j h + \sum_{i \in \mathcal{S}, i \neq j} \alpha_{0,i} q_{i,j} h + o(h). \tag{3.3}$$

This implies that over a small time interval of length h, we approximately observe the probability mass transitioning into state j from every state $i \neq j$ at a rate proportional to $q_{i,j}$, while probability mass leaves state j (and transitions to other states) at a rate proportional to $v_j = \sum_{k \neq j} q_{j,k}$.
Further, following a similar procedure as used to derive Equation (3.1), it is straightforward to have for $j \neq i$,

$$\mathbb{E}[\text{Transitions into } j \text{ in } [0, h] | X(0) = i] = q_{i,j} h + o(h),$$

and thus, the (expected) rate of transitions into state j given $X(0) = i$ as $h \to 0$ is

$$\lim_{h \to 0} \frac{\mathbb{E}[\text{Transitions into } j \text{ in } [0, h] | X(0) = i]}{h} = \lim_{h \to 0} \frac{q_{i,j} h + o(h)}{h} = q_{i,j}.$$

Hence, $q_{i,j}$ can be regarded as the instantaneous rate of transition into state j from state i.

3.3.1.3 Interpreting probability flows

Another interpretation of $q_{i,j}$ is as the rate of independent exponential timers. When the CTMC is in a non-absorbing state $i \in \mathcal{S}$, it initiates independent $\text{Exp}(q_{i,j})$ timers for all $j \in \mathcal{S}$. The CTMC transitions to the state corresponding to the shortest timer upon timer completion. If $q_{i,j} = 0$ for all $j \in \mathcal{S}$ ($v_i = 0$), state i is absorbing.
This interpretation aligns with the probabilistic behavior of CTMCs and is valuable when constructing the generator matrix.

3.3.1.4 Generator matrix and stationary distribution

Let us address the solution for the transition probability function matrix (TPFM) $P(t) = [P_{i,j}(t)]_{i,j\in\mathcal{S}}$, $t \geq 0$. Initially, note that $P_{i,j}(0) = \mathbb{P}(X(0) = j|X(0) = i) = \delta_{i,j}$, where $\delta_{i,j}$ denotes the Kronecker delta. This implies that $P(0) = I$, where I is the identity matrix. From Equations (3.1) and (3.2), we have

$$\lim_{h\to 0} \frac{P_{i,j}(h)}{h} = \lim_{h\to 0} \frac{q_{i,j}h + o(h)}{h} = q_{i,j},$$

and

$$\lim_{h\to 0} \frac{P_{i,j}(h) - 1}{h} = \lim_{h\to 0} \frac{-v_i h + o(h)}{h} = -v_i.$$

By employing the Chapman-Kolmogorov equations for CTMCs,

$$P_{i,j}(s+t) = \sum_{k\in S} P_{i,k}(s)P_{k,j}(t),$$

or in matrix form, $P(s+t) = P(s)P(t)$, we can derive the Kolmogorov Backward Equations (KBE):

$$\frac{d}{dt}P(t) = \lim_{h\to 0} \frac{P(t+h) - P(t)}{h} = \lim_{h\to 0} \frac{(P(h) - I)P(t)}{h} = QP(t), \tag{3.4}$$

and the Kolmogorov Forward Equations (KFE):

$$\frac{d}{dt}P(t) = \lim_{h\to 0} \frac{P(t+h) - P(t)}{h} = \lim_{h\to 0} \frac{P(t)(P(h) - I)}{h} = P(t)Q, \tag{3.5}$$

where the matrix Q is recognized as the generator (or infinitesimal generator) of $\{X(t), t \geq 0\}$, defined from [55] as:

Definition 3.2. *If $\{X(t), t \geq 0\}$ is a CTMC with transition probability matrix $P(t)$, the matrix Q is the infinitesimal generator matrix of $\{X(t), t \geq 0\}$ if*

$$Q = \lim_{h\to 0} \frac{P(h) - I}{h} = \begin{bmatrix} -v_0 & q_{0,1} & q_{0,2} & \cdots \\ q_{1,0} & -v_1 & q_{1,2} & \cdots \\ q_{2,0} & q_{2,1} & -v_2 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

From our understanding of the generator matrix Q, we can conclude that the duration of a sojourn at state i follows an exponential distribution with a rate of $-Q_{i,i} = v_i$. Additionally, the transition probability from state i to state j is $-Q_{i,j}/Q_{i,i} = q_{i,j}/v_i = p_{i,j}$. It is also noteworthy that the row sums of Q are all zero, as

$$\sum_{j\neq i} q_{i,j} = \sum_{j\neq i} v_i p_{i,j} = v_i \sum_{j\neq i} p_{i,j} = v_i.$$

Remark 3.3. *Equations (3.4) and (3.5) present differential equations whose solutions yield the transition probability matrix (TPM). We can verify that $P(t) = e^{tQ}$ fulfills both equations, with e^{tQ} denoting the matrix exponential function, defined as*

$$e^{tQ} = I + tQ + \frac{t^2}{2!}Q^2 + \ldots = \sum_{n=0}^{\infty} \frac{t^n}{n!} Q^n,$$

which converges to the standard Taylor series expansion of an exponential function when Q is a scalar. We can now represent the marginal distribution of $\{X(t), t \geq 0\}$ as $\alpha_t = \alpha_0 P(t)$ for $t \geq 0$.

Next, from [55], we obtain the stationary (or steady-state) distribution of a CTMC and establish its connection with the generator Q.

Definition 3.4. *If $\{X(t), t \geq 0\}$ is a CTMC with TPFM $P(t)$, a probability distribution $\{\pi_i\}_{i \in \mathcal{S}}$, where $\pi_i \geq 0$ for all $i \in \mathcal{S}$, is referred to as a stationary distribution of $\{X(t), t \geq 0\}$ if the probability row vector $\pi = (\pi_0, \pi_1, \ldots)$ satisfies the following conditions:*

$$\begin{aligned} &\pi e' = 1 \quad \textit{(Normalization condition)} \\ &\pi = \pi P(t), \quad \forall t \geq 0 \quad \textit{(Stationary condition)} \end{aligned}$$

It's important to note that in general, the notation e' is used to denote matrix (or vector) transpose, and e represents a row vector of ones.The set of probabilities in this vector is termed 'stationary' because, by setting $\alpha_0 = \pi$, we achieve $\alpha_t = \alpha_0 P(t) = \pi P(t) = \pi$. This observation signifies that if, at any given moment, the marginal distribution of $\{X(t), t \geq 0\}$ equals the probability vector π, then the continuous-time Markov chain (CTMC) has reached a state of stability. In such a scenario, the marginal distribution remains constant indefinitely, ensuring $\mathbb{P}(X(t) = i) = \pi_i$.
Drawing upon the definition of Q, we notice that if $\mathcal{S}$ consists of a finite number of states, then:

$$\begin{aligned} \pi &= \pi P(h), \\ 0 &= \pi(P(h) - I), \\ 0 &= \lim_{h \to 0} \pi_i \frac{P(h) - P(0)}{h} = \pi Q, \end{aligned} \tag{3.6}$$

where 0 is a row vector of zeroes with suitable dimensions. This offers a more straightforward method for solving for π, as we can directly apply Equation (3.6) without first deriving its TPFM. For the countable state-space case, assuming that π satisfies $\pi Q = 0$ and letting $\alpha_0 = \pi$, applying Equation (3.4) reveals that

$$\frac{d}{dt}\alpha_t = \frac{d}{dt}\alpha_0 P(t) = \alpha_0 P'(t) = \alpha_0 Q P(t) = \pi Q P(t) = 0P(t) = 0,$$

indicating that due to the KBE, the marginal distribution of $\{X(t), t \geq 0\}$ remains unchanged over time, establishing that π is a stationary distribution.
In the context of a continuous-time Markov chain (CTMC) $\{X(t), t \geq 0\}$ with an infinitesimal generator matrix Q, each element $q_{i,j}$ can be understood as the rate associated with an independent exponential timer. This interpretation offers practical insights when constructing the generator for a given model. Specifically:

- When $q_{i,j} = 0$, it implies that the corresponding timer takes on an infinite value with certainty.
- Upon entering a sojourn time at a non-absorbing state $i \in \mathcal{S}$, the CTMC initiates independent exponential timers with rates $\text{Exp}(q_{i,j})$ for all $j \in \mathcal{S}$.
- Upon the completion of the first timer, the CTMC concludes its visit to state i and switches to the state associated with the shortest timer duration, denoted as $j \neq i$. This process repeats iteratively.
- As long as there exists at least one $q_{i,j} > 0$, timers with $q_{i,j} = 0$ can effectively be disregarded, as they will never have the shortest duration.
- In cases where $q_{i,j} = 0$ for all $j \in \mathcal{S}$ (i.e., $v_i = 0$), the visit to state i never ends, as no timer ever reaches completion. Such a state is referred to as an absorbing state.

This alternative perspective provides valuable insights for the practical construction of the generator matrix in various modeling scenarios.

3.4 Quasi Birth-and-Death processes (QBD)

This section explores an $M/M/1$ queue, which are often called "quasi birth-death processes. This is based on Vuuren's Thesis [172]. The matrix-geometric method is used for analyzing the equilibrium behavior of $M/M/1$-type model.
Consider a Markov process with a state space composed of two distinct components: the boundary states denoted as $(0, j)$ for j ranging from 0 to n, and an infinitely extending set of states represented as (i, j) where i ranges from 1 to ∞, and j ranges from 0 to m.
These states are meticulously ordered in lexicographic fashion, encompassing sequences like $(0, 0), (0, 1), \ldots, (0, n), (1, 0), \ldots, (1, m), (2, 0), \ldots, (2, m), \ldots$.

Specifically, we designate the collection of boundary states as level 0 and the states within $\{(i,0),(i,1),\ldots,(i,n)\}$ for $i \geq 1$ as level i. It's worth noting that the number of states at level 0 may differ from those at higher levels, a characteristic common in many scenarios.
For this state space partitioning, we assume that the generator matrix Q takes the form:

$$Q = \begin{pmatrix} B_{00} & B_{01} & 0 & 0 & 0 & \ldots \\ B_{10} & A_1 & A_0 & 0 & 0 & \ldots \\ 0 & A_2 & A_1 & A_0 & 0 & \ldots \\ 0 & 0 & A_2 & A_1 & A_0 & \\ \vdots & \vdots & & \ddots & \ddots & \ddots \end{pmatrix},$$

where B_{00} is an $(n+1) \times (n+1)$ matrix, B_{01} is an $(n+1) \times (m+1)$ matrix, B_{10} is an $(m+1) \times (n+1)$ matrix, and A_0, A_1, and A_2 are square matrices of dimension $(m+1)$. Notably, $A_0 + A_1 + A_2$ forms a generator that characterizes the behavior of the Markov process Q solely in the vertical j-direction.
It's worth highlighting that the border of the Markov process can comprise multiple levels, rather than being confined to a single level as described previously.
Henceforth, we assume that the Markov process Q is irreducible, and the generator $A_0 + A_1 + A_2$ possesses precisely one communicating class. For the stability of Q, we assert the following result:

Theorem 3.5. *[131] The Markov process Q is ergodic (stable) if and only if*

$$\pi A_0 e < \pi A_2 e,$$

where e is the column vector of ones and $\pi = (\pi_0, \pi_1, \ldots, \pi_m)$ is the equilibrium distribution of the Markov process with generator $A_0 + A_1 + A_2$, so

$$\pi(A_0 + A_1 + A_2) = 0, \qquad \pi e = 1.$$

Condition $\pi A_0 e < \pi A_2 e$ holds an intuitive interpretation—the term $\pi A_0 e$ signifies the mean drift from level i to level $i+1$, while $\pi A_2 e$ represents the mean drift from level $i+1$ to level i. Hence, the process remains stable if the drift to the left surpasses that to the right (see the $M/M/1$ model, where rightward drift is λ and leftward drift is μ). Condition $\pi A_0 e < \pi A_2 e$ is widely recognized as Neuts' mean drift condition.

3.4.1 Matrix-geometric method

We highlight key results of MAM. It is important to note that for well exploration of matrix-geometric solutions, readers are encouraged to refer to

[131, 132]. In this section, only some of the main results are provided. Assuming the ergodicity of Markov process Q, we can elegantly express the equilibrium probability vectors π_i in matrix-geometric form:

$$\pi_i = (p(i,0), p(i,1), \dots, p(i,m)) = \pi_1 R^{i-1}, \quad i = 1, 2, \dots,$$

Here, the matrix R represents the "minimal nonnegative solution" to the matrix-quadratic equation:

$$A_0 + RA_1 + R^2 A_2 = 0. \tag{3.7}$$

It's noteworthy that any other nonnegative solution, denoted as $\tilde{R}$, of the above matrix equation satisfies $R \leq \tilde{R}$. The matrix R, often referred to as the rate matrix of Markov process Q, possesses a spectral radius less than one (implying that $I - R$ is invertible).
The rate matrix R also holds an intriguing and practical probabilistic interpretation. The element R_{jk} denotes the expected time spent in state $(i+1, k)$ before the first return to level i. This is expressed in time units of $-1/(A_1)_{jj}$, given the initial state (i, j). Notably, $-1/(A_1)_{jj}$ represents the expected time spent in state (i, j) with $i > 0$. From this interpretation, it follows that zero rows in A_0 correspond to zero rows in R. The equilibrium equations for the probability vectors π_0 and π_1 are as follows:

$$\begin{aligned} \pi_0 B_{00} + \pi_1 B_{10} &= 0, \\ \pi_0 B_{01} + \pi_1 A_1 + \pi_2 A_2 &= 0. \end{aligned}$$

Substituting $\pi_2 = \pi_1 R$, we obtain the boundary equations for π_0 and π_1 as follows:

$$\begin{aligned} \pi_0 B_{00} + \pi_1 B_{10} &= 0, \\ \pi_0 B_{01} + \pi_1 B_{11} + \pi_1 R A_2 &= 0. \end{aligned}$$

To uniquely determine π_0 and π_1, we need the normalization condition:

$$\sum_{i=0}^{\infty} \pi_i e = \pi_0 e + \pi_1 (I + R + R^2 + \dots) e = \pi_0 e + \pi_1 (I - R)^{-1} = 1.$$

For the computation of the matrix R, Equation (3.7) can be written as:

$$R = -(A_0 + R^2 A_2) A_1^{-1}.$$

It's worth noting that A_1 is indeed invertible, given that A_1 serves as a transient generator. We can solve the above fixed-point equation through successive substitutions, leading to the iteration:

$$R_{k+1} = -(A_0 + R_k^2 A_2) A_1^{-1}, \quad k = 0, 1, 2, \dots,$$

starting with $R_0 = 0$. Importantly, it can be demonstrated that as k tends to infinity, R_k converges to R. While this represents a straightforward scheme for computing R, the literature offers more advanced and efficient algorithms for this purpose, as detailed in [101].

Chapter 4

A matrix geometric solution of a multi-server queue with waiting servers and customers' impatience under variant working vacation and vacation interruption

4.1 Introduction

Queueing theory addresses one of life's most infuriating experiences; waiting. Queueing is very common in diverse areas, such as industry, emergency services, military logistics, finance, telecommunication systems, computer systems, and so on. This subject has attracted many researchers' attention [61, 62, 63].

Vacation queues have gained a particular focus since Levy and Yechiali [108] because of their excellent applications in various real-life problems, including manufacturing/production, inventory systems, computers and communication systems, and so on. Eminent surveys on these models can be found in [48, 154, 159, 82] and the references therein.

The concept of working vacation (WV) policy at which the server continues providing service at a lower rate during the vacation period has been introduced by Servi and Finn [145]. Over the last years, a great variety of queueing models with working vacations in different context has been done, for a detailed overview on the theme, the readers may refer to [12, 18, 73, 180, 158, 17, 74, 47, 187].

Nevertheless, we often come across the cases where the vacation may be interrupted, like for instance when number of customers reaches a predetermined value. Here, the interruption of the vacation avoids significant waiting costs for customers. This concept was initiated by Li and Tian [111] and Li *et al.* [113]. Since then, many studies have been provided on the subject (cf. [14, 197, 114, 49, 107, 160, 124]).

Customers' impatience has a very bad impact on different real-life systems including telecommunication, manufacturing and production systems. Working vacation queues with impatient customers have been well studied. The behavior of customers' impatience in working vacation queueing models have been extensively analysed. Prominent research papers can be found in [191, 144, 54, 32, 170, 52].

Moreover, impatience behavior in queueing models with variant of multiple vacation where the server can take a determined number of sequential vacations if there is no customers present in the queue at the end of a vacation have been considerably investigated (e.g., [81, 190, 104, 105]).

In different practical contexts, the server waits for a while before taking a break once the system gets empty. This frequently happens while considering the human behavior as a server. This topic has been thoroughly investigated (e.g., [134, 6, 46, 31, 152, 33]).

In practice, multi-server queues with station vacation (the servers, all together, synchronously go on vacation) and server vacation (the servers individually take vacations) are more applicable than single server queueing models. However, the analysis of these systems appears to be limited due to their complexity [115, 76, 123, 29, 184, 169, 24, 35, 25, 26, 70].

In recent decades, a large literature has been done on queueing models with impatience behavior of customers, nevertheless, no research work to date has investigated a variant working vacation queueing system with multiple servers, vacation interruption, waiting servers, balking and reneging. The present queueing model finds a powerful application in call centers (see Subsection 4.2.1).

The main goals of the current study are:

- To establish the stationary analysis of the suggested queueing system using the matrix geometric method;
- To derive different system characteristics.
- To develop a cost model in order to define the optimum service rates in both regular working period and working vacation period that optimize the total expected cost using the direct search method.

The remainder of this chapter is structured as follows. In Section 4.2, we describe the queueing model. In Section 4.3, we use the matrix geometric method to obtain the stationary probabilities of the system . In Section 4.4, the system performance measures as well as the expected cost function per unit time are developed . Section 4.5 provides some special cases of our study. Section 4.6 presents numerical examples for a sensitivity analysis and uses a direct search method to minimize the cost function, subject to the equilibrium condition. In 4.7 section, some conclusions are drawn.

4.2 The model

Consider a $M/M/c$ queueing system with K-kind of variant working vacations, vacation interruption, waiting servers, balking, and reneging:

- Customers arrive at the system in accordance to a Poisson process with rate λ.
- The service times during regular busy state are considered to be an i.i.d exponential random variables (r.v) with rate μ. The service discipline is FCFS.
- Whenever the system is emptied (the regular working period is ended), the servers stay idle before going on vacation (waiting servers), this period follows an exponential distribution with parameter ω.
- At a vacation completion, if no customers are present in the queue, the servers are allowed to take other vacations of shorter durations until the number of working vacations reached the maximum (defined by K-vacations), then the system returns to the regular working state, waiting for new customers. Type-j, $j = \overline{0, K-1}$ working vacation times are supposed to be i.i.d random variables that follow exponential distributions with parameter ϕ_j, where $\phi_j > \phi_{j-1}$.
- During the vacation time, incoming customers have the possibility for being serviced. Here, the service times are assumed to be an i.i.d exponential random variables with rate ν such that $\nu < \mu$. Within this period, when the service is completed, if there are some customers in the queue, the servers can stop (interrupt) the vacation under Bernoulli's rule and turn to the normal working state with probability β' or remain in the vacation state with probability $\beta = 1 - \beta'$. It should be noted that the service during the vacation can be offered only to the first arrival.
- If on arrival, the customer finds some servers idle, he will be directly served. Otherwise, the arrivals may join the queue with probability θ, or decide to balk with a complementary probability $\theta' = 1 - \theta$.
- During the working vacation time, the customers may get impatient and abandon the system (renege) if their services are not yet accomplished. Here, the impatience times are supposed to be i.i.d random variables that follow exponential distributions with rate ξ.

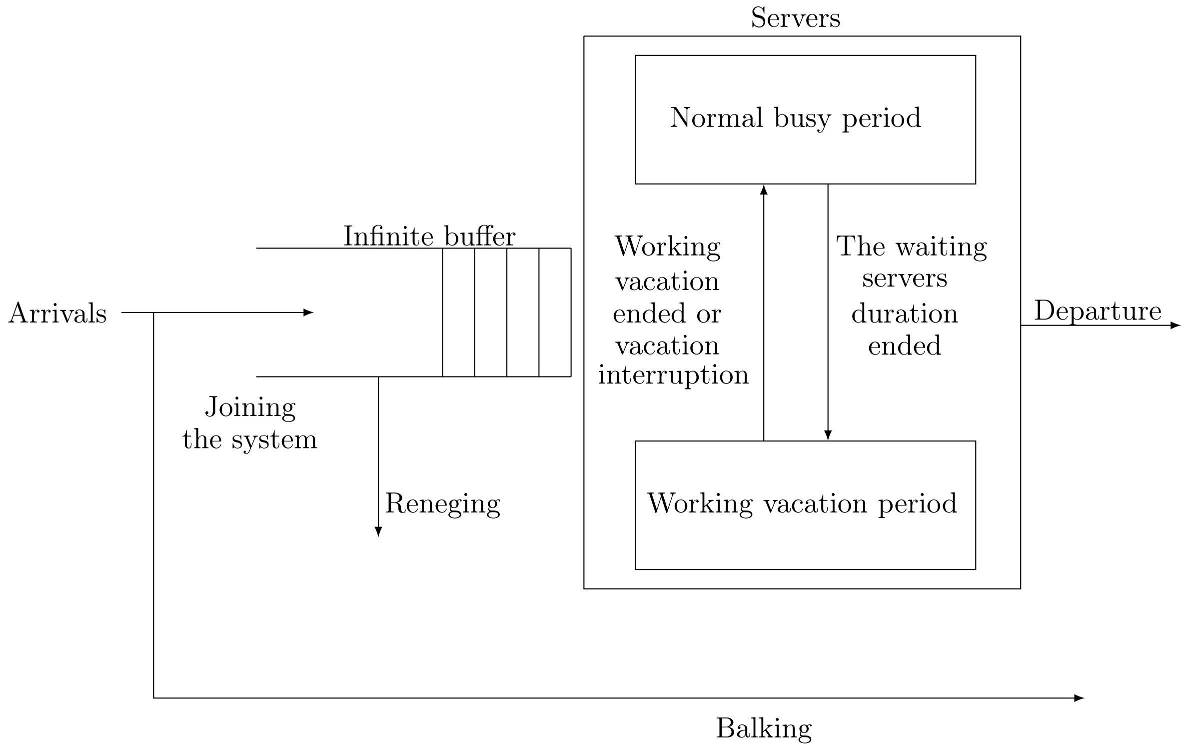

Figure 4.1: Schematic representation of the queueing system.

- The above variables (the inter-arrival times, service times, waiting server times, working vacation times, are impatience times) are mutually independent.

The considered model is schematically depicted in the Figure 4.1.

4.2.1 Practical application

The proposed queueing model has a potential application in call centers, where the calls can reach the call center after traversing through different intermediate routers. If the agents are unoccupied, a call request is immediately processed. Otherwise, they will be waiting in a buffer to be serviced according to first-come-first-serve (FCFS) policy. When there is no calls in the system, before going synchronously 'as a group' on vacation, they will wait a random period of time (waiting servers duration). During the vacation period, the agents can deal with the new calls (if any) at the slower rate to economize the cost (working vacation period). During this period, at the end of each service, the agents check if there are new calls in the system and decide whether or not to return from their vacation, whether it is over or not (vacation interruption). In addition, if there is no calls request at a vacation completion, the agents starts a finite number of working vacations (variant working vacation). Otherwise, the agents begin a new regular busy period. During both periods, the call may balk based on the queue length. Further, during working vacation, a waiting call in the system may become impatient and quit the system after a long wait (reneging).

4.3 Analysis of the model

The behaviour of our queueing system is described by two-dimensional infinite state continuous-time Markov chain $\{(S(t); L(t)); t \geq 0\}$ with state space $\Omega = \{(j, n) : n \geq 0, j = \overline{0, K}\}$, where $L(t)$ stands for the number of customers in the system and $S(t)$ specifies the state of the servers at time t, where

$$S(t) = \begin{cases} j, & \text{the system is on } (j+1)^{th} \text{ WV at time } t,\ j = \overline{0, K-1}; \\ K, & \text{the system is on regular working state at time } t. \end{cases}$$

Let $P_{j,n} = \lim\limits_{t \to \infty} P\{S(t) = j, L(t) = n\}$, $n \geq 0, j = \overline{0, K}$ be the system state probabilities of the process $\{(S(t); L(t)), t \geq 0\}$. Before proceeding with the analysis of the queueing system let us consider the following notations that are necessary for the rest of the article:

$$\chi_{0,n} = \begin{cases} \nu + \xi, & n = 1, \\ n(\beta\nu + \xi), & 2 \leq n \leq c-1, \\ c\beta\nu + n\xi, & n \geq c, \end{cases} \qquad \chi_{1,n} = \begin{cases} n\mu, & 1 \leq n \leq c-1, \\ \\ c\mu, & n \geq c, \end{cases}$$

and

$$\alpha_n = \begin{cases} 0, & n = 0, 1 \\ n(1-\beta)\nu, & 2 \leq n \leq c-1, \\ c(1-\beta)\nu, & n \geq c. \end{cases}$$

The transition diagram is illustrated in Figure 4.2.

4.3.1 Matrix-geometric method

In this section, we apply the matrix geometric method to obtain the stationary probabilities. According to [131], the infinitesimal generator $\mathbf{Q}$ for the process could be given as:

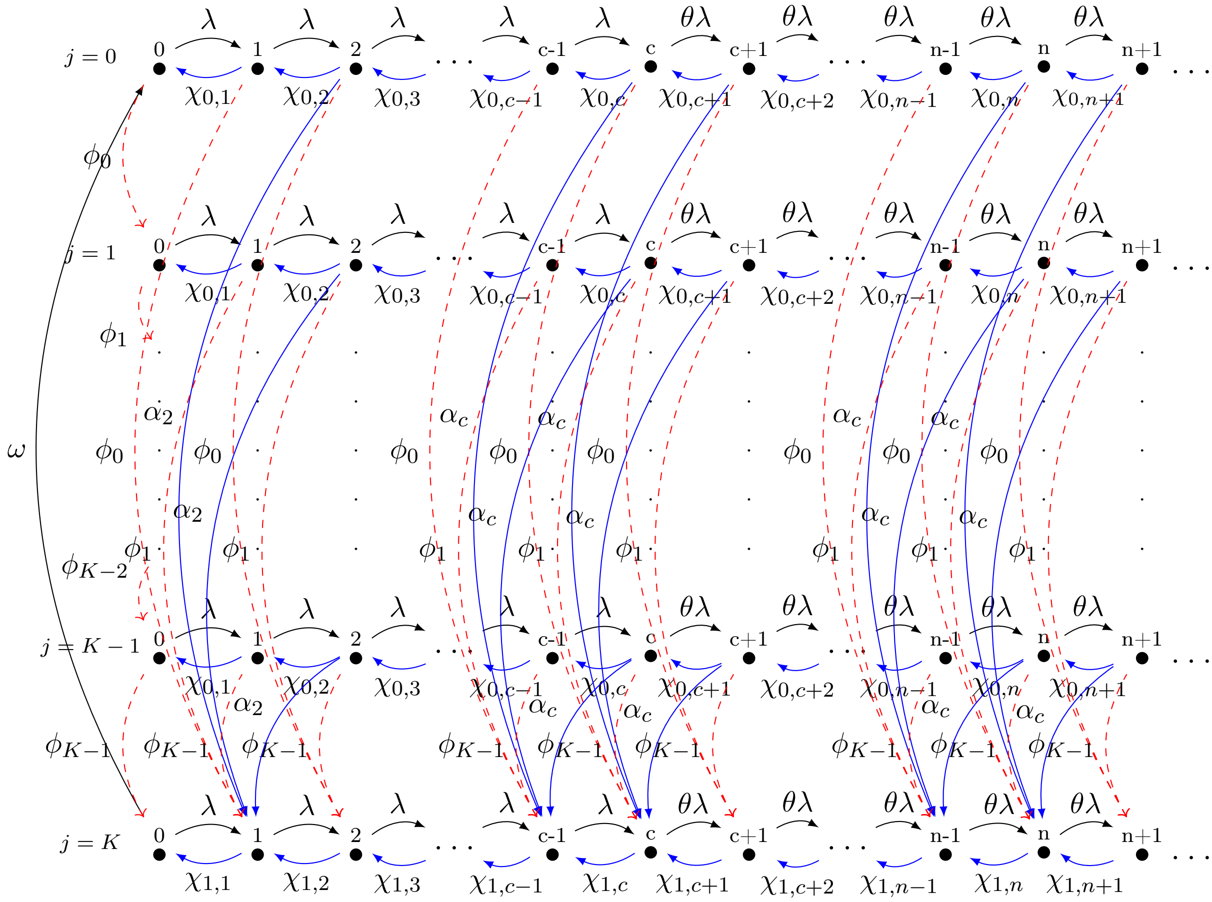

Figure 4.2: State-transition-rate diagram

$$
\mathbf{Q} = \begin{pmatrix}
\mathbf{A}_0 & \mathbf{C}_0 & & & & & & & & & \\
\mathbf{B}_1 & \mathbf{A}_1 & \mathbf{C}_0 & & & & & & & & \\
& \mathbf{B}_2 & \mathbf{A}_2 & \mathbf{C}_0 & & & & & & & \\
& & \ddots & \ddots & \ddots & & & & & & \\
& & & \mathbf{B}_{c-1} & \mathbf{A}_{c-1} & \mathbf{C}_0 & & & & & \\
& & & & \mathbf{B}_c & \mathbf{A}_c & \mathbf{C}_1 & & & & \\
& & & & & \mathbf{B}_{c+1} & \mathbf{A}_{c+1} & \mathbf{C}_1 & & & \\
& & & & & & \ddots & \ddots & \ddots & & \\
& & & & & & & \mathbf{B}_N & \mathbf{A}_N & \mathbf{C}_1 & \\
& & & & & & & & \mathbf{B}_N & \mathbf{A}_N & \mathbf{C}_1 \\
& & & & & & & & & \ddots & \ddots
\end{pmatrix},
$$

where N is a large enough number so that when the number of customers $n \geq N$, we approximate the matrices $\mathbf{A}_n$ and $\mathbf{B}_n$ by $\mathbf{A}_N$ and $\mathbf{B}_N$, respectively.

$$
\mathbf{C}_0 = \begin{pmatrix}
\lambda & & & & & & \\
& \lambda & & & & & \\
& & \ddots & & & & \\
& & & \ddots & & & \\
& & & & \ddots & & \\
& & & & & \lambda & \\
& & & & & & \lambda
\end{pmatrix}_{K+1 \times K+1},
$$

$$
\mathbf{C}_1 = \begin{pmatrix}
\theta\lambda & & & & & & \\
& \theta\lambda & & & & & \\
& & \ddots & & & & \\
& & & \ddots & & & \\
& & & & \ddots & & \\
& & & & & \theta\lambda & \\
& & & & & & \theta\lambda
\end{pmatrix}_{K+1 \times K+1},
$$

$$
\mathbf{A}_0 = \begin{pmatrix}
-(\lambda+\phi_0) & \phi_0 & & & & \\
& -(\lambda+\phi_1) & \phi_1 & & & \\
& & \ddots & \ddots & & \\
& & & \ddots & \ddots & \\
& & & & -(\lambda+\phi_{K-1}) & \phi_{K-1} \\
\omega & & & & & -(\lambda+\omega)
\end{pmatrix}_{K+1 \times K+1},
$$

$$\mathbf{B}_1 = \begin{pmatrix} \chi_{0,1} & & & & & & \\ & \chi_{0,1} & & & & & \\ & & \ddots & & & & \\ & & & \ddots & & & \\ & & & & \ddots & & \\ & & & & & \chi_{0,1} & \\ & & & & & & \mu \end{pmatrix}_{K+1\times K+1},$$

$$\mathbf{B}_n = \begin{pmatrix} \chi_{0,n} & & & & & & n\beta'\nu \\ & \chi_{0,n} & & & & & n\beta'\nu \\ & & \ddots & & & & \vdots \\ & & & \ddots & & & \vdots \\ & & & & \ddots & & \vdots \\ & & & & & \chi_{0,n} & n\beta'\nu \\ & & & & & & n\mu \end{pmatrix}_{K+1\times K+1}, \; 2 \leq n \leq c-1,$$

$$\mathbf{B}_n = \begin{pmatrix} \chi_{0,n} & & & & & & c\beta'\nu \\ & \chi_{0,n} & & & & & c\beta'\nu \\ & & \ddots & & & & \vdots \\ & & & \ddots & & & \vdots \\ & & & & \ddots & & \vdots \\ & & & & & \chi_{0,n} & c\beta'\nu \\ & & & & & & c\mu \end{pmatrix}_{K+1\times K+1}, \; c \leq n \leq N-1,$$

$$\mathbf{A}_n = \begin{pmatrix} -\Theta_0 & & & & \phi_0 \\ & -\Theta_1 & & & \phi_1 \\ & & \ddots & & \vdots \\ & & & \ddots & \vdots \\ & & & -\Theta_{K-1} & \phi_{K-1} \\ & & & & -(\lambda+n\mu) \end{pmatrix}_{K+1\times K+1}, \; 1 \leq n \leq c-1,$$

where $\Theta_j = (\lambda + \phi_j + \alpha_n + \chi_{0,n})$, $0 \leq j \leq K-1$,

$$\mathbf{A}_n = \begin{pmatrix} -\delta_0 & & & & \phi_0 \\ & -\delta_1 & & & \phi_1 \\ & & \ddots & & \vdots \\ & & & \ddots & \vdots \\ & & & -\delta_{K-1} & \phi_{K-1} \\ & & & & -(\theta\lambda+n\mu) \end{pmatrix}_{K+1\times K+1}, \; c \leq n \leq N-1,$$

where $\delta_j = (\theta\lambda + \phi_j + \chi_{0,n} + c\beta'\nu)$, $0 \leq j \leq K-1$,

$$\mathbf{B}_n = \begin{pmatrix} \chi_{0,N} & & & & & c\beta'\nu \\ & \chi_{0,N} & & & & c\beta'\nu \\ & & \ddots & & & \vdots \\ & & & \ddots & & \vdots \\ & & & & \ddots & \vdots \\ & & & & \chi_{0,N} & c\beta'\nu \\ & & & & & c\mu \end{pmatrix}_{K+1\times K+1}, \; n \geq N,$$

$$\text{and } \mathbf{A}_n = \begin{pmatrix} -\Gamma_0 & & & & & \phi_0 \\ & -\Gamma_1 & & & & \phi_1 \\ & & \ddots & & & \vdots \\ & & & \ddots & & \vdots \\ & & & & -\Gamma_{K-1} & \phi_{K-1} \\ & & & & & -(\theta\lambda + c\mu) \end{pmatrix}_{K+1\times K+1}, \; n \geq N,$$

where $\Gamma_j = (\theta\lambda + \phi_j + \chi_{0,N} + c\beta'\nu)$, $0 \leq j \leq K-1$.
Based on Neuts [131], the approximated system is stable and the steady-state probability vector exists iff $\mathbf{YC}_1\mathbf{e} < \mathbf{YB}_N\mathbf{e}$, where $\mathbf{Y}$ is an invariant probability of the matrix $\psi = \mathbf{B}_N + \mathbf{A}_N + \mathbf{C}_1$. $\mathbf{Y}$ satisfies the equations $Y\psi = 0$ and $\mathbf{Ye}_n = 1$, where $\mathbf{e}_n$ is the column vector of appropriate dimension n with all elements equal to one.

Under the stability condition, the stationary probability vector $\mathbf{\Pi}$ of the generator $\mathbf{Q}$ exists, satisfying the balance equation $\mathbf{\Pi Q} = \mathbf{0}$ and $\mathbf{\Pi e}_n = 1$, where $\mathbf{0}$ is the row vector with all elements equal to zero. The vector $\mathbf{\Pi}$ partitioned as $\mathbf{\Pi} = [\mathbf{\Pi}_0, \mathbf{\Pi}_1, \mathbf{\Pi}_2, ...]$, where $\mathbf{\Pi}_n = [P_{0,n}, P_{1,n}, P_{2,n}, ..., P_{K,n}]$.

Clearly, when the stability condition is fulfilled, the sub-vectors of $\mathbf{\Pi}$, relating to various levels satisfy

$$\mathbf{\Pi}_n = \mathbf{\Pi}_N \mathbf{R}^{n-N}, \; n \geq N, \tag{4.1}$$

where the matrix $\mathbf{R}$ is the minimal non-negative solution of the matrix quadratic equation

$$\mathbf{C}_1 + \mathbf{RA}_N + \mathbf{R}^2\mathbf{B}_N = \mathbf{0}. \tag{4.2}$$

In fact, the QBD process is positive recurrent iff the spectral radius $Sp(\mathbf{R}) < 1$. However, it is quite complicated and tedious to define the explicit expression of the matrix $\mathbf{R}$ by resolving equation (4.2). Neuts [131] developed an iterative algorithm for numerically computing $\mathbf{R}$. Starting with initial iteration $\mathbf{R}_0 = 0$, we can compute the successive approximation

$$\mathbf{R}_{n+1} = -(\mathbf{C}_1 + \mathbf{R}_n^2\mathbf{B}_N)(\mathbf{A}_N)^{-1}, n \geq 0.$$

From equation $\mathbf{\Pi Q} = \mathbf{0}$, the governing system of difference equations can be given as

$$\mathbf{\Pi}_0\mathbf{A}_0 + \mathbf{\Pi}_1\mathbf{B}_1 = 0, \tag{4.3}$$
$$\mathbf{\Pi}_{n-1}\mathbf{C}_0 + \mathbf{\Pi}_n\mathbf{A}_n + \mathbf{\Pi}_{n+1}\mathbf{B}_{n+1} = 0, \quad 1 \le n \le c \tag{4.4}$$
$$\mathbf{\Pi}_{n-1}\mathbf{C}_1 + \mathbf{\Pi}_n\mathbf{A}_n + \mathbf{\Pi}_{n+1}\mathbf{B}_{n+1} = 0, \quad c+1 \le n \le N-1, \tag{4.5}$$
$$\mathbf{\Pi}_{n-1}\mathbf{C}_1 + \mathbf{\Pi}_n\mathbf{A}_N + \mathbf{\Pi}_{n+1}\mathbf{B}_N = 0, \quad n \ge N, \tag{4.6}$$

and the normalizing condition

$$\sum_{n=0}^{\infty} \mathbf{\Pi}_n\mathbf{e} \quad = \quad 1. \tag{4.7}$$

From equations (4.3) to (4.6), after some mathematical manipulations, we obtain

$$\mathbf{\Pi}_{n-1} = \mathbf{\Pi}_n\varphi_n, \quad 1 \le n \le N, \tag{4.8}$$
$$\mathbf{\Pi}_N[\varphi_N\mathbf{C}_1 + \mathbf{A}_N + \mathbf{R}\mathbf{B}_N] = \mathbf{0}, \tag{4.9}$$

where

$$\varphi_1 = -\mathbf{B}_1(\mathbf{A}_0^{-1}), \quad \varphi_n = -\mathbf{B}_n(\mathbf{A}_{n-1} + \varphi_{n-1}\mathbf{C}_0)^{-1}, 2 \le n \le c+1,$$
$$\varphi_n = -\mathbf{B}_n(\mathbf{A}_{n-1} + \varphi_{n-1}\mathbf{C}_1)^{-1}, c+2 \le n \le N.$$

Using equations (4.7) and (4.8), we obtain

$$\mathbf{\Pi}_N\left[\sum_{n=1}^{N}\prod_{i=N}^{n}\varphi_i + (\mathbf{I}-\mathbf{R})^{-1}\right]\mathbf{e} \quad = \quad 1. \tag{4.10}$$

By solving equations (4.9) and (4.10), we find $\mathbf{\Pi}_N$. Then, We employ equations (4.1) and (4.8) to obtain $\mathbf{\Pi}_n$ for $n \ge 0$.

4.4 Performance measures and cost model

4.4.1 Performance measures

- The expected number of customers in the system during WV:

$$E[L_{wv}] = \sum_{j=0}^{K-1}\sum_{n=1}^{\infty} nP_{j,n}.$$

- The expected number of customers in the system during regular working period:

$$E[L_K] = \sum_{n=1}^{\infty} nP_{K,n}.$$

- The expected number of customers in the system:

$$E[L] = E[L_{wv}] + E[L_K].$$

- The servers remain idle during regular working period with probability

$$P[I_b] = P_{K,0}.$$

- The servers remain idle during WV period with probability

$$P_{idle} = \sum_{j=0}^{K-1} P_{j,0}.$$

- The probability that the servers are in WV period is given as

$$P_{wv} = \sum_{j=0}^{K-1} \sum_{n=0}^{\infty} P_{j,n}.$$

- The probability that the servers are busy during regular working state is as follows:

$$P_{busy} = 1 - P_{K,0} - \sum_{j=0}^{K-1} P_{j,0}.$$

- The average rate of reneging is

$$R_{ren} = \xi E[L_{wv}].$$

- Throughput is given as

$$T_P = \mu \sum_{n=1}^{c-1} n P_{K,n} + c\mu \sum_{n=c}^{\infty} P_{K,n} + \nu \sum_{j=0}^{K-1} \sum_{n=1}^{\infty} P_{j,n}$$

4.4.2 Cost model

We develop a cost model to analyze the optimization of the cost function of the model. The total expected cost function per unit time is as:
$F[\mu, \nu] = c_l \, E[L] + c_b \, (\mu + \nu) P_{busy} + c_r \, R_{ren} + c_i \, P[I_b] + C_\mu \, \mu,$
where
$C_l \equiv$ Cost per unit time per customer present in the system ,
$C_b \equiv$ Cost per unit time when the servers are busy,
$C_r \equiv$ Cost per unit time when a customer reneges,
$C_i \equiv$ Cost per unit time when the servers are idle during busy period,
$C_\mu \equiv$ Fixed service purchase cost per unit during busy period.
The cost minimization problem is illustrated mathematically as:

$$F[\mu^*, \nu] = \text{minimize}_\mu F[\mu, \nu],$$

$$F[\mu^*, \nu^*] = \text{minimize}_\nu F[\mu^*, \nu].$$

4.5 Some special case

In this section, we present some important particular cases of our queueing model.

- For both $K \to \infty$, and $K = 1$, if $c = 1$, $\nu = 0$, $\phi_i = \phi$, $i = \overline{0, K-1}$, $\omega = 0$, $\beta = 1$, and $\theta = 1$, then our model reduces to the models investigated in Altman and Yechiali [4].

- If $c = 1$, $\nu = 0$, $\phi_i = \phi$, $i = \overline{0, K-1}$, $\omega = 0$, $\xi = 0$, $\beta = 1$, and $\theta = 1$, our system is reduced with that studied by Yue *et al.* [190].

- When $c = 1$, $\phi_i = \phi$, $i = \overline{0, K-1}$, $\omega = 0$, $\xi = 0$, and $\beta = 1$, our queueing model coincides with that examined by Vijaya Laxmi and Rajesh [104].

- When $\phi_i = \phi$, $i = \overline{0, K-1}$, $\omega = 0$, $\theta = 1$, and $\beta = 1$, and the customers may be impatient before the service begins, then the queueing model presented in the current chapter match with that given by Vijaya Laxmi and Kassahun [169].

4.6 Numerical illustrations

Numerical computations were carried out in this section using Mathematica software, and the results are provided in the form of graphs given below. Unless their values are indicated in the appropriate places, the model parameters are assumed to be $\lambda = 1.0$, $\mu = 5.0$, $\beta = 0.6$, $\nu = 3.0$, $\theta = 0.5$, $\omega = 0.7$, $\xi = 0.8$, $\phi[0] = 0.5$, $\phi_i = \phi_{i-1} + 0.1$, $1 \leq i \leq K-1$. Cost parameters are taken as $c_l = 25$, $c_b = 20$, $c_r = 10$, $c_i = 6$, and $c_\mu = 5$.

Figure 4.3 shows the effect of arrival rate λ on P_{busy} and probability $P_{K,0}$, for different values of service rate in busy period μ. As λ increases, P_{busy} and $P_{K,0}$ increases and decreases, respectively. This is because, an increase in customers inflow into the system increases the probability of busy servers resulting in a decrease in server idleness. Further, for a fixed λ, contrary trend is observed as μ increases, which is true. The point of intersection of curves is indicated by the value of λ at which P_{busy} and $P_{K,0}$ are the maximum and minimum, respectively.

The impact of ν on $E[L]$, for different number of working vacations K, is shown in Figure 4.4. The figure shows that for a fixed K, the increase in ν decreases $E[L]$, which is obvious. Moreover, for a constant ν, $E[L]$ shows an opposite trend with the increase of K because of the slower service rates during the vacations. Also, $E[L]$ is observed smaller for $K = 1$.

In Figure 4.5, we discuss the effect of λ on $E[L]$ for two scenarios; queueing model without balking and without reneging. We notice that when there is balking but no reneging, the system size is larger, and smaller when there

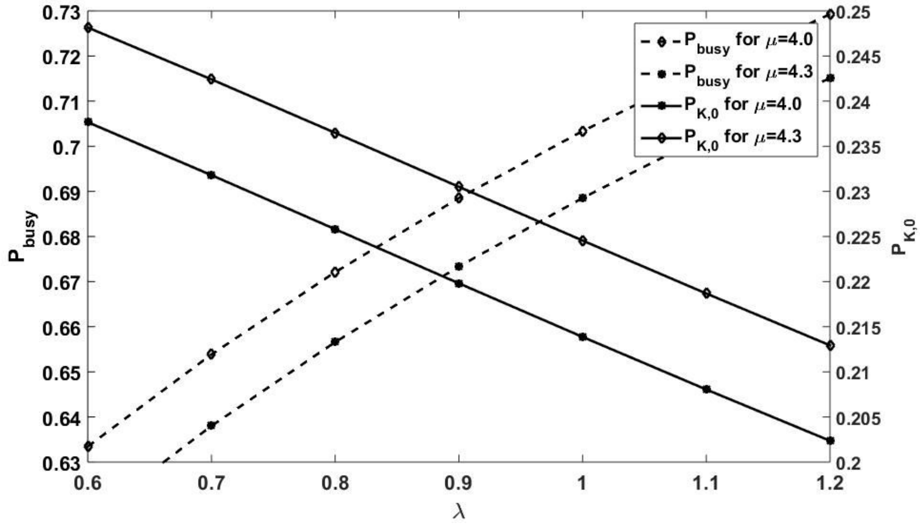

Figure 4.3: Effect of λ on P_{busy} and $P_{K,0}$ for different μ

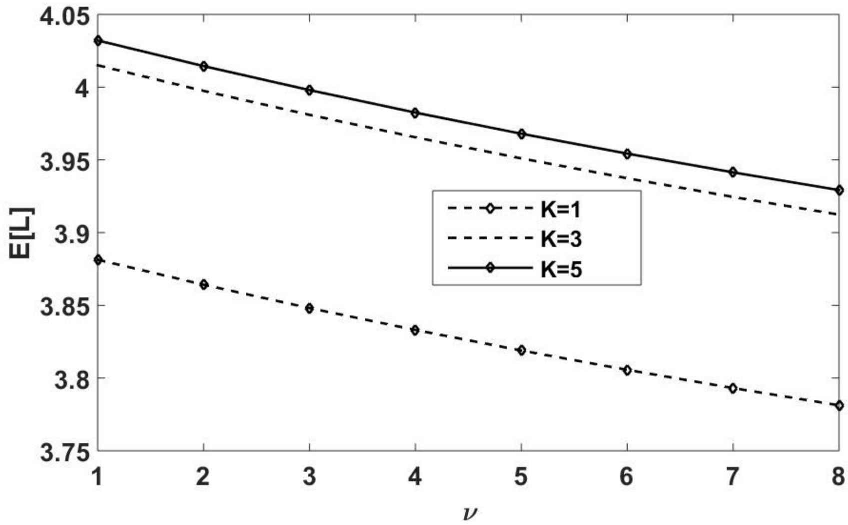

Figure 4.4: Effect of ν on $E[L]$ for different K

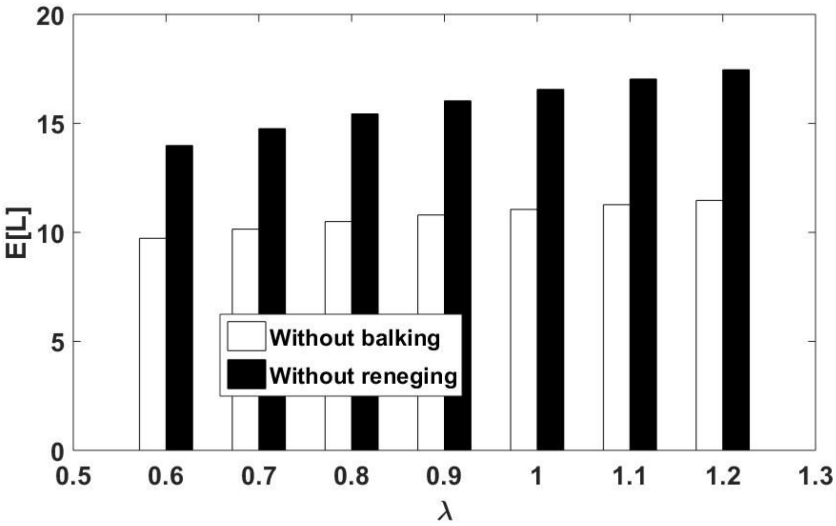

Figure 4.5: Effect of λ on $E[L]$

is reneging but no balking. This demonstrates that for this system reneging constraint has a negative impact than balking factor.

Figure 4.6 depicts the impact of number of servers c on throughput of the system T_p with and without reneging. As we see, the increase in c increases T_p. Also, for a fixed c, T_p is observed higher in the absence of reneging because of a longer queue.

The impact of λ on R_{ren} is shown, for various vacation interruption probabilities β', in Figure 4.7. Initially, average reneging rate of the customer is high when there is no vacation interruption ($\beta' = 0$). Further, as β' increases, R_{ren} tends to decrease. This is due to as β' increases, the servers spend more in busy period and also service rates are higher during regular busy period.

In Figure 4.8, we compare $E[L]$ in three cases $\theta' <=> \xi$ with respect to λ. This figure reveals that the system length is larger when $\theta' > \xi$ and smaller for $\theta' < \xi$. Hence, for the multi-server systems with balking and reneging of customers, in order to sustain the system properly, reneging rate should be maintained smaller than balking probability.

Using direct search method, the effect of λ, ξ, and c on optimum cost $F[\mu^*, \nu^*]$, optimum service rates, during busy period μ^*, and working vacation period ν^*, are shown in Tables 4.1-4.3, respectively.

- As arrival rate λ increases, the optimum service rates μ^* and ν^* decrease and the minimum cost $F[\mu^*, \nu^*]$ increases, which is necessary in order to maintain the stability of the system.

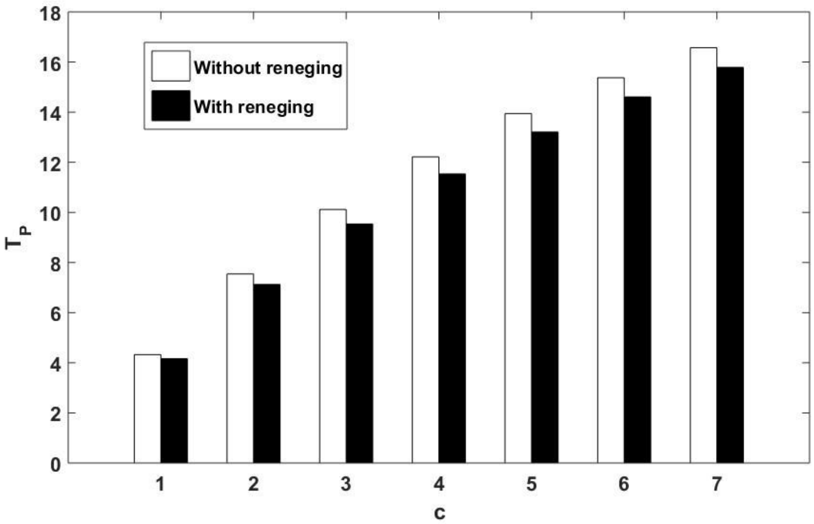

Figure 4.6: Effect of c on T_P

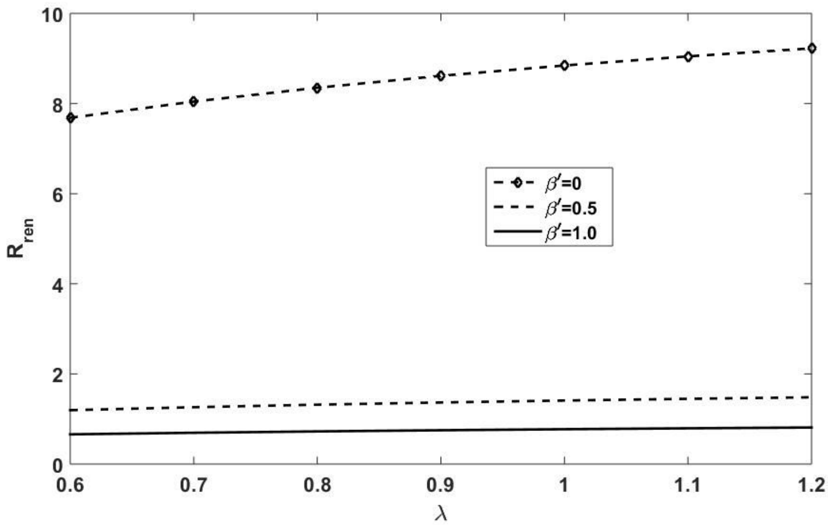

Figure 4.7: Effect of λ on R_{ren} for different β'

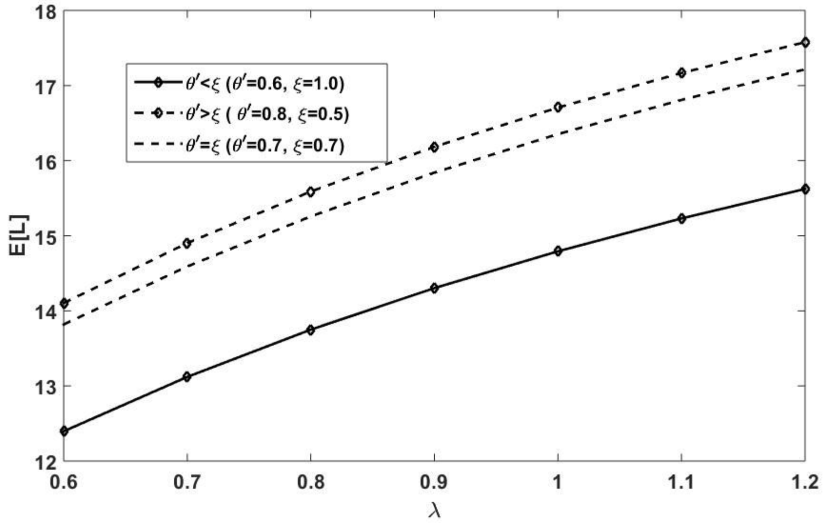

Figure 4.8: Effect of λ on $E[L]$ for different θ' and ξ

Table 4.1: Effect of λ on cost function.

λ	(μ^*, ν)	$F[\mu^*, \nu]$	(μ^*, ν^*)	$F[\mu^*, \nu^*]$
0.8	(7.5,3.0)	471.331	(7.5,2.2)	463.783
1.0	(6.8,3.0)	509.298	(6.8,1.9)	501.366
1.2	(6.0,3.0)	539.570	(6.0,1.7)	528.748
1.4	(5.0,3.0)	563.339	(5.0,1.4)	546.111

Table 4.2: Effect of ξ on cost function.

ξ	(μ^*, ν)	$F[\mu^*, \nu]$	(μ^*, ν^*)	$F[\mu^*, \nu^*]$
0.5	(7.2,3.0)	524.532	(7.2,2.4)	520.176
0.7	(6.9,3.0)	514.101	(6.9,2.0)	507.029
0.9	(6.7,3.0)	504.835	(6.7,1.9)	496.425
1.1	(6.5,3.0)	496.982	(6.5,1.7)	488.481

Table 4.3: Cost function vs. c

c	(μ^*, ν)	$F[\mu^*, \nu]$	(μ^*, ν^*)	$F[\mu^*, \nu^*]$
3	(5.1,3.0)	547.081	(5.1,1.3)	525.976
5	(7.7,3.0)	473.907	(7.7,2.6)	470.265
7	(8.4,3.0)	416.053	(8.4,3.7)	411.996
9	8.5,3.0	372.56	(8.5,4.3)	367.004

- Further, the optimum service rates and the minimum cost decrease with the increase of reneging rate ξ. This agrees with the fact that to attract the reneged customers in the system there should be some decrease in the minimum cost.
- However, the minimum cost decreases and the optimum service rates grow with the increase of number of servers c.

4.7 Conclusion

In this chapter, we investigated a multi-server queue with K- variant working vacations, vacation interruption, waiting servers and customers' impatience. The stationary solution of the queueing system is established. Different performance measures are derived. In addition, cost optimization along with numerical results are presented. Our results show that

1. An increase in the number of the servers increases the throughput of the system.
2. The average reneging rate lowers as the probability of the vacation interruption increases.
3. For the better maintenance of the system, reneging rate should be smaller than balking probability.
4. The minimum cost and optimal service rates are reduced when the reneging rate rises.

The method used in this chapter can be applied to study different Markovian models, such as $Geo^X/G/c$ and $GI^X/Geo/1$ queues with variant working vacations, Bernoulli-schedule vacation interruption and impatient customers.

Chapter 5

$M/M/c/DWV$ queueing model with Bernoulli schedule working vacation interruption, waiting servers, and impatience as a model of a call center

5.1 Introduction

The subject of vacation queues in queueing theory has been, for an extended period of time, an important area of research. This is due to its diverse practical applications like manufacturing and production systems, telephone systems, mail systems, file transfer services, and so on. Readers can refer to the renowned monographs and books afforded by [108, 48, 154, 159] to learn more about the topic, and for the latest developments in the field, see the recent research papers [58, 83, 128, 1, 129].

In recent times, researchers have suggested an alternate service mechanism with different service rates in the vacation period. During this time, the server continues serving new arrivals at a slower rate. The infinite-space single-server Markovian queueing model with working vacation (WV) was first initiated by [145]. Then, the result was extended by [180] and [12] to the $M/G/1$ queue and $GI/M/1$ respectively. Later, [8] introduced $M/G/1/WV$ retrial queue. An $M/G/1$ feedback retrial queueing model with breakdown and repair under multiple working vacation has been discussed in [136]. Moreover, during the working vacation duration, if some customers are present at the service completing moment, the server may discontinue the vacation state and come back to the normal busy period. This policy, termed as vacation interruption, was introduced in [49] and [50] while developing $M/G/1$ queue with single working vacation and $M/G/1$ retrial queueing model with working vacation and general retrial times, respectively. For more details on the theme, the readers may refer to [38, 80, 161, 143, 151].

Over the last few years, a special attention has been paid to the impatience behavior in queueing systems, especially in vacation queueing models. This fascinating aspect of the investigation revealed a substantial impact on queueing applications, particularly in manufacturing, and production systems, call centers, computer networks. This concept has been the focus of multiple researchers (cf. [189, 144, 166, 28, 168, 27, 24, 26]).

5.1.1 Motivation of the current research work

The benefits and contributions of this article are along these lines:

- **Suggested queueing model:** This article addresses an $M/M/c/DWV$ with Bernoulli interruption schedule, waiting servers, along with the presence of impatient customers. It is appropriate to note that the differentiated vacation (DV) approach has been lately introduced by [71, 72]. In these queues, the system considers two types of vacations; long and short vacations. The longer vacation period is taken as soon as the server has terminated serving at least one customer (type-1 vacation) while the shorter vacation is taken instantly after the server has just come back from the preceding vacation to find the system empty (type-2 vacation). Great attention has been devoted to this novel notion given its large applicability in real-time situations like call centers, power save/sleep mode for energy-efficient utilization in modern mobile technologies, hospital emergency room operation, gas stations, etc. Readers interested in this subject can refer to [33, 152, 59, 162].
  The present research comprises concurrently the following concepts:
  - Infinite-capacity multi-server Markovian queues
  - Differentiated vacations
  - Waiting servers
  - Working vacations
  - Vacation interruption
  - Bernoulli interruption schedule
  - Balking and reneging
- **Theoretical results:** We build a Markovian process of the considered queueing system. Then, by employing the matrix geometric method, the steady-state probabilities are obtained. Various performance measures in a steady state are also derived.

- **Numerical representations:** We carry out the sensitivity analysis to illustrate the impact of the system parameters on the performance characteristics, which can provide insight to the system managers to manage the operation status of the system and reduce the congestion problem. In addition, we develop a cost function to help the system managers or decision-makers regulate the system economically.

This chapter is organized as follows. In Section 5.2, a detailed description of the mathematical model under consideration is given. Section 5.3 is dedicated to the steady-state analysis by the matrix geometric method. In Section 5.4, various important performance indices of our model are given. In Section 5.5, numerical results for sensitivity analysis are provided, where a cost-revenue-profit study is provided. At last, concluding remarks and suggestions for future work are given in Section 5.6.

5.2 Model description and practical application

5.2.1 Model description

We consider an $M/M/c/DWV$ queueing system with waiting servers, balking, and reneging under the Bernoulli interruption schedule. The basic assumption of this queueing model is as:

- **The arrival process:** Customers enter the system according to a Poisson process with rate λ.

- **The service process during regular busy periods:** Customers are served following First Come First Served 'FCFS' discipline. The service time during the normal busy period follows an exponential distribution with rate μ.

- **Waiting server process:** If no arrival occurs during the regular busy period, the servers, before going on a vacation, keep waiting (inactive) for a possible new arrival for a random time. The waiting server period follows an exponential distribution with rate η.

- **Vacation process:** After the waiting period expires, the system (the servers together) leaves for an initial type-1 vacation exponentially distributed with parameter γ_1. Upon return from the initial type-1 vacation, if the servers do not find any customers waiting in the queue, they leave for the second type-2 vacation of shorter duration which follows exponential distributions with parameter γ_2. Otherwise, they return to the regular state. When there are no customers left in the waiting queue after completing type-II of vacation, the servers come back to the busy period to await new customers.

- **The service process during vacation busy periods:** During vacation periods, customers are serviced at slower rates. The service times during these periods follow exponential distributions at rates ν_1, (during type-1 vacation) and ν_2, (during type-2 vacation).

- **Bernoulli interruption schedule:** At a service completion moment during the working vacation periods, if the servers notice that some customers are present in the queue, then the working vacation state can be stopped with probability β' or kept going on with a complementary probability $\beta = 1 - \beta'$. It should be noted that the service during this time can be only addressed to the primary customer.

- **Balking:** On arrival, during regular busy periods (resp. either type-1 or type-2 working vacations), if a customer observes that some servers are busy and some others are idle, he joins the queue and will be serviced promptly. Otherwise, he may decide to enter the system with a certain probability θ_1 (resp. θ_2) or balk with a complementary probability $\theta_1' = 1 - \theta_1$ (resp. $\theta_2' = 1 - \theta_2$).

- **Reneging process:** In type-1 and type-2 working vacation periods, incoming customers may lose their patience and quit the system if they are not yet served. The impatience times follow exponential distributions with respective parameters ξ_1 and ξ_2.

- The above variables (of customers as well as servers status) are mutually independent.

The general structure of the suggested queueing model described above is illustrated in Figure 5.1.

5.2.2 Practical justification of the proposed queueing model

The proposed queueing model finds practical application in call centers, which serve as great examples of its functionality. In a typical call center, there is a finite number of customer service agents available to handle incoming customer interactions through phone calls and live chat. When all agents are occupied with ongoing calls, incoming calls are queued in an infinite queue.

During periods of inactivity when no incoming calls are received, agents enter an idle state and may remain inactive for a random duration. During this idle time, they perform maintenance tasks to enhance computer performance, referred to as a type-I vacation. Following a type-I vacation, if the system remains empty, agents can opt for a shorter type-2 vacation. Meanwhile, new incoming calls are processed at a reduced rate. At the end of a type-2 vacation, agents return to the regular busy state and await new arrivals. When service is completed during both types of vacations, agents

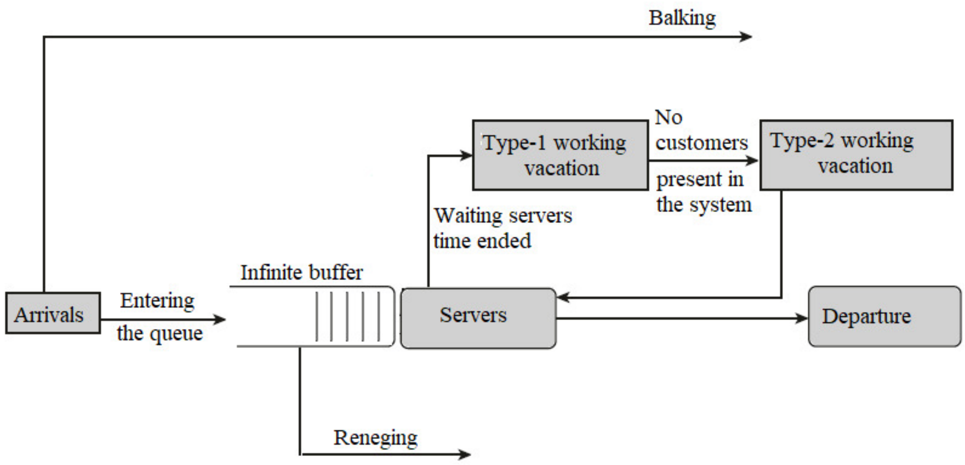

Figure 5.1: An $M/M/c/DWV$ with waiting servers, balking, and reneging

may choose to continue handling incoming calls or transition to the regular busy period, providing service at a higher rate.

Upon arrival, if a call observes that all lines are occupied, it enters the system and occupies a free line. Calls unable to access a free line have two options: they may join a queue with a certain probability or immediately abandon the system, a behavior known as balking.

During vacation periods, if at least one server is available, incoming calls receive immediate service. Otherwise, they must wait in a queue. During this waiting time, calls may become impatient and choose to abandon (renege) the system, even if they are currently being served.

5.3 Analysis of the steady-state probability distribution

Let $L(t)$ represent the number of customers in the system at time t, and let $S(t)$ denote the status of the servers at time t. For the mathematical representation of the proposed model at an instant t, we consider the following states of the system based on the status of the servers:

$$S(t) = \begin{cases} 0, & \text{for regular busy period;} \\ 1, & \text{for type-1 working vacation;} \\ 2, & \text{for type-2 working vacation.} \end{cases}$$

The bi-variate process $\{(L(t); S(t)); t \geq 0\}$ represents two-dimensional infinite state Markov chain in continuous time with state space:

$$\Omega = \{(j, n) : j = 0, 1, 2, n \geq 0\}.$$

Let

$$\pi_{j,n} = \lim_{t \to \infty} \mathbb{P}\{S(t) = j, L(t) = n\}, \;\; j = 0, 1, 2, \;\; n \geq 0$$

denote the system steady-state probabilities of the process $\{(S(t); L(t)), t \geq 0\}$. The transition diagram is illustrated in Figure 5.2.

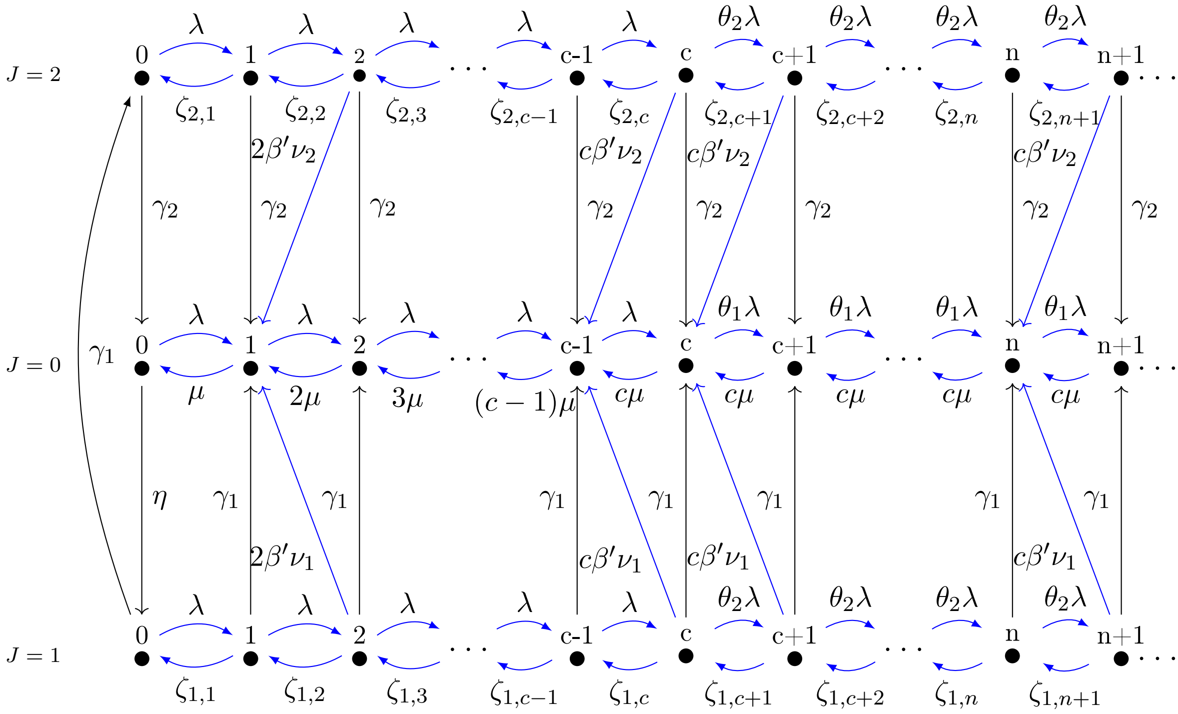

Figure 5.2: State transition rate diagram.

1. If $J(t) = 0$, the normal busy period:

$$
\begin{aligned}
(\lambda+\eta)\pi_{0,0} &= \mu\pi_{0,1} + \gamma_2\pi_{2,0}, \;\; n = 0, && (5.1)\\
(\lambda+\mu)\pi_{0,1} &= \lambda\pi_{0,0} + 2\mu\pi_{0,2} + \gamma_1\pi_{1,1} + \gamma_2\pi_{2,1} + \\
&\quad 2\beta'\nu_1\pi_{1,2} + 2\beta'\nu_2\pi_{2,2}, \;\; n = 1, && (5.2)\\
(\lambda+n\mu)\pi_{0,n} &= \lambda\pi_{0,n-1} + (n+1)\mu\pi_{0,n+1} + \gamma_1\pi_{1,n} + \\
&\quad \gamma_2\pi_{2,n} + (n+1)\beta'\nu_1\pi_{1,n+1} \\
&\quad +(n+1)\beta'\nu_2\pi_{2,n+1}, \;\; 2 \le n \le c-1, && (5.3)\\
(\theta_1\lambda+c\mu)\pi_{0,n} &= \lambda\pi_{0,n-1} + c\mu\pi_{0,n+1} + \gamma_1\pi_{1,n} + \gamma_2\pi_{2,n} \\
&\quad +c\beta'\nu_1\pi_{1,n+1} + c\beta'\nu_2\pi_{2,n+1}, \\
&\quad n = c, && (5.4)\\
(\theta_1\lambda+c\mu)\pi_{0,n} &= \theta_1\lambda\pi_{0,n-1} + c\mu\pi_{0,n+1} + \gamma_1\pi_{1,n} + \gamma_2\pi_{2,n} \\
&\quad +c\beta'\nu_1\pi_{1,n+1} + c\beta'\nu_2\pi_{2,n+1}, \\
&\quad n > c. && (5.5)
\end{aligned}
$$

2. If $J(t) = 1$, type-1 working vacation period:

$$
\begin{aligned}
(\lambda+\gamma_1)\pi_{1,0} &= \eta\pi_{0,0} + (\nu_1+\xi_1)\pi_{1,1}, \;\; n = 0, && (5.6)\\
(\lambda+\zeta_{1,1}+\gamma_1)\pi_{1,1} &= \lambda\pi_{1,0} + \zeta_{1,2}\pi_{1,2}, \;\; n = 1, && (5.7)\\
(\lambda+n(\nu_1+\xi_1)+\gamma_1)\pi_{1,n} &= \lambda\pi_{1,n-1} + \zeta_{1,n+1}\pi_{1,n+1}, \;\; 2 \le n \le c-1, && (5.8)\\
(\theta_2\lambda+c\nu_1+n\xi_1+\gamma_1)\pi_{1,n} &= \lambda\pi_{1,n-1} + \zeta_{1,n+1}\pi_{1,n+1}, \;\; n = c, && (5.9)\\
(\theta_2\lambda+c\nu_1+n\xi_1+\gamma_1)\pi_{1,n} &= \theta_2\lambda\pi_{1,n-1} + \zeta_{1,n+1}\pi_{1,n+1}, \quad n > c. && (5.10)
\end{aligned}
$$

3. If $J(t) = 2$, type-2 working vacation period:

$$
\begin{aligned}
(\lambda+\gamma_2)\pi_{2,0} &= \gamma_1\pi_{1,0} + (\nu_2+\xi_2)\pi_{2,1}, \;\; n = 0, && (5.11)\\
(\lambda+\zeta_{2,1}+\gamma_2)\pi_{2,1} &= \lambda\pi_{2,0} + \zeta_{2,2}\pi_{2,2}, \;\; n = 1, && (5.12)\\
(\lambda+n(\nu_2+\xi_2)+\gamma_2)\pi_{2,n} &= \lambda\pi_{2,n-1} + \zeta_{2,n+1}\pi_{2,n+1}, \;\; 2 \le n \le c-1, && (5.13)\\
(\theta_2\lambda+c\nu_2+n\xi_2+\gamma_2)\pi_{2,n} &= \lambda\pi_{2,n-1} + \zeta_{2,n+1}\pi_{2,n+1}, \;\; n = c, && (5.14)\\
(\theta_2\lambda+c\nu_2+n\xi_2+\gamma_2)\pi_{2,n} &= \theta_2\lambda\pi_{2,n-1} + \zeta_{2,n+1}\pi_{2,n+1}, \quad n > c, && (5.15)
\end{aligned}
$$

where for $j = 1, 2$ we have:

$$
\zeta_{j,n} = \begin{cases} \nu_j + \xi_j, & n = 1, \\ n(\beta\nu_j + \xi_j), & 2 \le n \le c-1, \\ c\beta\nu_j + n\xi_j, & n \ge c, \end{cases}
$$

The normalizing condition is defined as:

$$
\sum_{n=0}^{\infty} \pi_{0,n} + \sum_{n=0}^{\infty} \pi_{1,n} + \sum_{n=0}^{\infty} \pi_{2,n} = 1.
$$

5.3.1 Infinitesimal generator of the process

Equations (5.1)-(5.15) can be given in the following matrix form:

$$\begin{cases} \mathbf{\Pi Q} = \mathbf{0}, \\ \sum_{n=0}^{\infty} \mathbf{\Pi}_n \mathbf{e}_n = 1, \end{cases} \tag{5.16}$$

where $\mathbf{e}_n$ is a column vector with appropriate dimension and all elements equal to 1. Let $\mathbf{\Pi} = (\mathbf{\Pi}_0, \mathbf{\Pi}_1, ..., \mathbf{\Pi}_{N-1}, \mathbf{\Pi}_N, \mathbf{\Pi}_{N+1}, ...)$ be the steady-state probability vector of the transition rate matrix $\mathbf{Q}$, and $\mathbf{\Pi_n}$ is defined as:

$$\mathbf{\Pi}_n = (\pi_{0,n}, \pi_{1,n}, \pi_{2,n}).$$

The transition rate matrix $\mathbf{Q}$ can be written as:

$$\mathbf{Q} = \begin{pmatrix} \mathbf{A}_0 & \mathbf{C}_0 & & & & & & & & & \\ \mathbf{B}_1 & \mathbf{A}_1 & \mathbf{C}_0 & & & & & & & & \\ & \mathbf{B}_2 & \mathbf{A}_2 & \mathbf{C}_0 & & & & & & & \\ & & \ddots & \ddots & \ddots & & & & & & \\ & & & \mathbf{B}_{c-1} & \mathbf{A}_{c-1} & \mathbf{C}_0 & & & & & \\ & & & & \mathbf{B}_c & \mathbf{A}_c & \mathbf{C}_1 & & & & \\ & & & & & \mathbf{B}_{c+1} & \mathbf{A}_{c+1} & \mathbf{C}_1 & & & \\ & & & & & & \ddots & \ddots & \ddots & & \\ & & & & & & & \mathbf{B}_N & \mathbf{A}_N & \mathbf{C}_1 & \\ & & & & & & & & \mathbf{B}_N & \mathbf{A}_N & \mathbf{C}_1 \\ & & & & & & & & & \ddots & \ddots \end{pmatrix},$$

where N is a sufficiently large number such that when the number of customers $n \geq N$, we approximate the matrices $\mathbf{A}_n$ and $\mathbf{B}_n$ by $\mathbf{A}_N$ and $\mathbf{B}_N$, respectively. This is a way of truncating the infinite state space of the Markov chain to a finite one, which makes the computation of the stationary distribution easier. This is a common assumption in queueing theory, where for large queue lengths, the service rates are often assumed to be practically independent of the queue lengths. The value of N can be chosen so that the losses due to balking and reneging is small. Due to the intrinsic nature of the system, the only choice available for studying N is through algorithmic methods. While a number of approaches are available for determining the cut-off point, N, the one that seems to perform well (with respect to approximating the system performance measures) is to start with a small value and increase it until the difference between two consecutive solutions is below a certain tolerance level.

Each sub-matrix of the matrix $\mathbf{Q}$ is done as:

$$\mathbf{A}_0 = \begin{pmatrix} -(\lambda+\eta) & \eta & 0 \\ 0 & -(\lambda+\gamma_1) & \gamma_1 \\ \gamma_2 & 0 & -(\lambda+\gamma_2) \end{pmatrix}, \quad \mathbf{C}_0 = \begin{pmatrix} \lambda & & \\ & \lambda & \\ & & \lambda \end{pmatrix},$$

$$\mathbf{B}_1 = \begin{pmatrix} \mu & 0 & 0 \\ 0 & \nu_1+\xi_1 & 0 \\ 0 & 0 & \nu_2+\xi_2 \end{pmatrix}, \quad \mathbf{C}_1 = \begin{pmatrix} \theta_1\lambda & & \\ & \theta_2\lambda & \\ & & \theta_2\lambda \end{pmatrix},$$

$$\mathbf{A}_1 = \begin{pmatrix} -(\lambda+\mu) & & \\ \gamma_1 & -(\lambda+\nu_1+\xi_1+\gamma_1) & \\ \gamma_2 & & -(\lambda+\nu_2+\xi_2+\gamma_2) \end{pmatrix},$$

$$\mathbf{B}_n = \begin{pmatrix} n\mu & & \\ n\beta'\nu_1 & n(\beta\nu_1+\xi_1) & \\ n\beta'\nu_2 & & n(\beta\nu_2+\xi_2) \end{pmatrix}, \quad 2 \leqslant n \leqslant c-1$$

$$\mathbf{B}_n = \begin{pmatrix} c\mu & & \\ c\beta'\nu_1 & c\beta\nu_1+n\xi_1 & \\ c\beta'\nu_2 & & c\beta\nu_2+n\xi_2 \end{pmatrix}, \quad c \leqslant n \leqslant N-1$$

$$\mathbf{B}_n = \begin{pmatrix} c\mu & & \\ c\beta'\nu_1 & c\beta\nu_1+N\xi_1 & \\ c\beta'\nu_2 & & c\beta\nu_2+N\xi_2 \end{pmatrix}, \quad n \geq N$$

$$\mathbf{A}_n = \begin{pmatrix} -(\lambda+n\mu) & & \\ \gamma_1 & -(\lambda+n(\nu_1+\xi_1)+\gamma_1) & \\ \gamma_2 & & -(\lambda+n(\nu_2+\xi_2)+\gamma_2) \end{pmatrix}, \quad 2 \leqslant n \leqslant c-1$$

$$\mathbf{A}_n = \begin{pmatrix} -(\theta_1\lambda+c\mu) & & \\ \gamma_1 & -(\theta_2\lambda+c\nu_1+n\xi_1+\gamma_1) & \\ \gamma_2 & & -(\theta_2\lambda+c\nu_2+n\xi_2+\gamma_2) \end{pmatrix}, \quad c \leqslant n \leqslant N-1$$

$$\mathbf{A}_n = \begin{pmatrix} -(\theta_1\lambda+c\mu) & & \\ \gamma_1 & -(\theta_2\lambda+c\nu_1+N\xi_1+\gamma_1) & \\ \gamma_2 & & -(\theta_2\lambda+c\nu_2+N\xi_2+\gamma_2) \end{pmatrix}, \quad n \geq N$$

From [131], we know that $\mathbf{Q}$ is seen as the infinitesimal generator of the process $\{(S(t);L(t)), t \geq 0\}$.

5.3.2 Equilibrium condition of the system

Theorem 5.1. *The system maintains stable if and only if we have*

$$\theta_1\lambda < c\mu. \tag{5.17}$$

Proof. Based on [131], the approximated system is stable and the steady-state probability vector exists if and only if

$$\mathbf{x}\mathbf{C_1}\mathbf{e}_n < \mathbf{x}\mathbf{B}_N\mathbf{e}_n, \tag{5.18}$$

where $\mathbf{x} = [x_1, x_2, x_3]$ is the invariant probability vector of the matrix:

$$\mathbf{F} = \mathbf{B}_N + \mathbf{A}_N + \mathbf{C}_1,$$

and $\mathbf{e}_n$ denotes a column vector with size n with all elements equal to one. Further, $\mathbf{x}$ satisfies

$$\begin{cases} \mathbf{xF} = \mathbf{0}, \\ \mathbf{xe}_n = 1. \end{cases} \tag{5.19}$$

Solving the above two equations, we get

$$\mathbf{x} = [x_1, x_2, x_3] = [1, 0, 0].$$

Then, by substituting $\mathbf{x}$, $\mathbf{e}_n$, $\mathbf{C}_1$, and $\mathbf{B}_N$ into Equation (5.18), we find the stability condition (5.17). □

5.3.3 Stationary probability distribution

We use the matrix-geometric method to calculate the stationary probability distribution of the number of customers in the system. The stationary probability vectors are given by

$$\mathbf{\Pi}_n = \mathbf{\Pi}_N \mathbf{R}^{n-N}, \ \ n \geq N, \tag{5.20}$$

where the rate matrix $\mathbf{R}$ is the unique non-negative solution of the quadratic equation:

$$\mathbf{C}_1 + \mathbf{R}\mathbf{A}_N + \mathbf{R}^2\mathbf{B}_N = 0. \tag{5.21}$$

In fact, the QBD process is positive recurrent if and only if the spectral radius is less than one. Nevertheless, it is difficult to get an explicit expression of the rate matrix $\mathbf{R}$ by solving the nonlinear Equation (5.21). Based on [131] and [101], one can derive an approximate solution of $\mathbf{R}$ by considering the sequence $\{\mathbf{R}_n\}$ as:

$$\mathbf{R}_0 = 0, \ \text{ and } \ \mathbf{R}_{n+1} = (\mathbf{C}_1 + \mathbf{R}_n^2\mathbf{B}_N)(-\mathbf{A}_N)^{-1}, \ n \geq 0. \tag{5.22}$$

As the sequence $\{\mathbf{R}_n\}$ is monotone, by successive substitutions, we can evaluate $\mathbf{R}$ from Equation (5.24). In the iterative procedure for finding $\mathbf{R}$, we have used the convergence criteria $||\mathbf{R}_{n+1} - \mathbf{R}_n|| < \epsilon$, where ϵ is a pre-defined tolerance limit.

From Equation (5.16), $\mathbf{\Pi Q} = \mathbf{0}$, we have:

$$\begin{aligned}
\mathbf{\Pi}_0\mathbf{A}_0 + \mathbf{\Pi}_1\mathbf{B}_1 &= 0, && (5.23)\\
\mathbf{\Pi}_{n-1}\mathbf{C}_0 + \mathbf{\Pi}_n\mathbf{A}_n + \mathbf{\Pi}_{n+1}\mathbf{B}_{n+1} &= 0, \ 1 \leq n \leq c, && (5.24)\\
\mathbf{\Pi}_{n-1}\mathbf{C}_1 + \mathbf{\Pi}_n\mathbf{A}_n + \mathbf{\Pi}_{n+1}\mathbf{B}_{n+1} &= 0, \ c+1 \leq n \leq N-1, && (5.25)\\
\mathbf{\Pi}_{n-1}\mathbf{C}_1 + \mathbf{\Pi}_n\mathbf{A}_N + \mathbf{\Pi}_{n+1}\mathbf{B}_N &= 0, \ n \geq N, && (5.26)\\
& && (5.27)
\end{aligned}$$

and the normalizing condition

$$\sum_{n=0}^{\infty} \mathbf{\Pi}_n \mathbf{e}_n \quad = \quad 1. \tag{5.28}$$

From equations (5.23)-(5.26), after some mathematical manipulations, we find:

$$\mathbf{\Pi}_0 = \mathbf{\Pi}_1 \mathbf{B}_1 (-\mathbf{A}_0)^{-1} = \mathbf{\Pi}_1 \mathbf{\Phi}_1, \tag{5.29}$$
$$\mathbf{\Pi}_{n-1} = \mathbf{\Pi}_n B_n [-(\Phi_{n-1} C_0 + A_{n-1})]^{-1} = \Pi_n \Phi_n, \;\; 2 \leq n \leq c+1, \tag{5.30}$$
$$\mathbf{\Pi}_{n-1} = \mathbf{\Pi}_n \mathbf{B}_n [-(\mathbf{\Phi}_{n-1} \mathbf{C}_1 + \mathbf{A}_{n-1})]^{-1} = \mathbf{\Pi}_n \mathbf{\Phi}_n, \;\; c+2 \leq n \leq N, \tag{5.31}$$
$$\mathbf{\Pi}_N [\mathbf{\Phi}_N \mathbf{C}_1 + \mathbf{A}_N + \mathbf{R}\mathbf{B}_N] = 0, \;\; n \geq N. \tag{5.32}$$

From the normalizing condition (5.28) and Equations (5.29)-(5.31), we obtain

$$\sum_{n=0}^{\infty} \mathbf{\Pi}_n \mathbf{e}_n \quad = \quad \sum_{n=0}^{N-1} \mathbf{\Pi}_n \mathbf{e}_n + \sum_{n=N}^{\infty} \Pi_n \mathbf{e}_n \tag{5.33}$$
$$= \quad \mathbf{\Pi}_N \left(\sum_{n=1}^{N} \prod_{i=N}^{m} \Phi_i + (\mathbf{I} - \mathbf{R})^{-1} \right) \mathbf{e}_n = 1, \tag{5.34}$$

where $\Phi_1 = -\mathbf{B}_1 (\mathbf{A}_0)^{-1}$, and

$$\Phi_n = \begin{cases} -\mathbf{B}_n [(\Phi_{n-1} \mathbf{C}_0 + \mathbf{A}_{n-1})]^{-1}, & 2 \leq n \leq c+1; \\ -\mathbf{B}_n [(\Phi_{n-1} \mathbf{C}_1 + \mathbf{A}_{n-1})]^{-1}, & c+2 \leq n \leq N. \end{cases}$$

Finally, $\mathbf{\Pi}_N$ can be obtained by solving Equations (5.32)-(5.34), and $\mathbf{\Pi}_n$ will be determined by Equations (5.20), (5.29)-(5.31).

5.4 System performance measures

Some significant system performance measures of the $M/M/c/DWV$ with Bernoulli interruption schedule, waiting servers, balking and reneging are described in this Section.

– The mean number of customers in the system:

$$E(L) = \sum_{n=1}^{\infty} n(\pi_{0,n} + \pi_{1,n} + \pi_{2,n})$$

– The mean number of customers in the queue:

$$E(L)_q = \sum_{n=c}^{\infty} (n-c)(\pi_{0,n} + \pi_{1,n} + \pi_{2,n}).$$

– Average rate of balked customers:

$$B_r = \theta_1' \lambda \sum_{n=c}^{\infty} \pi_{0,n} + \theta_2' \lambda \sum_{n=c}^{\infty} (\pi_{1,n} + \pi_{2,n}).$$

– The average reneging rate:

$$R_r = \xi_1 \sum_{n=1}^{\infty} n\pi_{1,n} + \xi_2 \sum_{n=1}^{\infty} n\pi_{2,n}.$$

– The system remains in the regular working state with probability:

$$P_b = \sum_{n=0}^{\infty} \pi_{0,n}.$$

– The system remains in type-1 WV (resp. in type-2 WV) with respective probabilities:

$$P_{wv1} = \sum_{n=0}^{\infty} \pi_{1,n} \text{ and } P_{wv2} = \sum_{n=0}^{\infty} \pi_{2,n}.$$

– The system remains inactive in regular working state with probability:

$$P_{id} = \pi_{0,0}.$$

– The servers keep working during regular busy state with probability:

$$P_w = 1 - \pi_{0,0} - (P_{wv1} + P_{wv2}).$$

– Expected number of customers served per unit time:

$$ECS = \mu \sum_{n=1}^{c-1} n\pi_{0,n} + c\mu \sum_{n=c}^{\infty} \pi_{0,n} + \nu_1 \sum_{n=1}^{c-1} n\pi_{1,n} + c\nu_1 \sum_{n=c}^{\infty} \pi_{1,n} + \nu_2 \sum_{n=1}^{c-1} n\pi_{2,n} + c\nu_2 \sum_{n=c}^{\infty} \pi_{2,n}.$$

5.5 Special cases of our model

In this section, we show how our queueing model relates to some previous studies in the literature, where it has been assumed that the vacation process continues even if there are no customers in the queue after completing type-II of vacation. We identify the parameters and assumptions that make our model equivalent to some existing models.

1. Our queueing model coincides with the one in [152] when $c = 1$ (single server), $\theta_1 = \theta_2 = 1$ (no balking probability), $\xi_1 = \xi_2 = \xi$ (same reneging rate), $\nu_1 = \nu_2 = 0$ (no service rate during vacation period), and $\beta = 1$ (no vacation interruption).

2. Our queueing model is reduced to the one in [33] when $c = 1$, $\theta_1 = \theta_2 = \theta$, $\nu_1 = \nu_2 = \nu$, and the customers are impatient during busy period with an exponential distribution of rate 'ξ_0', and the system has a finite capacity of 'N'.

3. If we set $c = 1$, $\theta_1 = \theta_2 = \theta = 1$, $\nu_1 = \nu_2 = 0$, $\eta \to \infty$ (no waiting servers), $\beta = 1$, and assume that the system has a finite capacity of 'L', and there is no reneging ($\xi_1 = \xi_2 = 0$), then our queueing model coincides with the one discussed in [162].

4. Our queueing model is reduced to the one in [71] when $c = 1$, $\theta_1 = \theta_2 = \theta = 1$, $\nu_1 = \nu_2 = \nu = 0$, $\eta \to \infty$, $\beta = 1$, and there is no reneging ($\xi_1 = \xi_2 = 0$).

5.6 Numerical analysis

Employing Mathematica software, we present numerically and graphically, the impact of system parameters on various system performance measures and total cost, revenue, and profit functions.

5.6.1 Cost-revenue-profit modeling

The effect of the cost function as well as the revenue function, along with the profit function on the system parameters is discussed. Here, it's noteworthy that the cost structure and notation are in accordance with well-known results [153, 187]. Further, the arbitrary values of the parameters are chosen so that the steady state condition is verified. To develop a cost model, we define the governing unit cost elements as:

- C_l : Cost per unit time per customer present in the system: This represents the expenses incurred for each customer in the system at any given time.

- C_i : Cost per unit time when the servers are idle during busy period: This signifies the cost associated with servers idleness before going on vacation.

- C_w : Cost per service per unit time when the servers are working: This cost pertains to the service cost during active working periods when the servers are actively engaged in customer service.

- C_r : Cost per unit time when a customer reneges: It represents the cost attributed to customers who leave the system due to impatience or dissatisfaction.

- C_μ : Fixed service purchase cost per unit during regular busy period: This cost is associated with maintaining regular service during busy periods.
- C_{ν_i} : Fixed service purchase cost per unit during working vacation period: This cost relates to maintaining service during working vacation periods, which are characterized by slower service rates.

Using the definition of each cost element and its corresponding system characteristics, the total expected cost per unit time of the system is given by

$$TCF = C_l E(L) + C_i P_{id} + C_w P_w + C_r R_r + c\mu C_\mu + c(\nu_1 + \nu_2) C_{\nu_i}.$$

The total revenue function and the total profit functions are defined as

$$T_r = Revenue \times ECS; \quad T_p = T_r - TCF,$$

where $Revenue$ is the constant revenue generated through the effective throughput ECS and obviously, $T_p \geq 0$ indicates profit otherwise it is a loss system.

5.6.2 Parameter impact study

To study the qualitative aspects of the queueing model under consideration, numerical computations have been carried out and some of them are presented in the form of tables and graphs. For computational purpose, the values are taken as $N = 50, \epsilon = 10^-5$.

Table 5.1: Effect of λ on performance measures for $\eta = 0.6$, $\gamma_1 = 0.5$, $\gamma_2 = 0.7$, $\beta = 0.5$, $\theta_1 = 0.6$, $\theta_2 = 0.4$, $\mu = 4.0$, $\nu_1 = 3.5$, $\nu_2 = 2.5$, $\xi_1 = 0.5$, $\xi_2 = 0.3$, $c = 5$

λ	E[L]	P_{id}	P_b	B_r	R_r	P_{wv1}	P_{wv2}
0.2	5.31887	0.300952	0.53447	0.0152484	0.00360932	0.291116	0.174414
0.4	8.70304	0.277338	0.660057	0.0499025	0.00508897	0.224556	0.115387
0.6	11.0655	0.255515	0.74246	0.0951766	0.0055903	0.177813	0.0797271
0.8	12.8218	0.23534	0.799513	0.147048	0.00561573	0.143546	0.0569412
1.	14.1874	0.216778	0.840583	0.203392	0.00540804	0.117658	0.0417589
1.2	15.2853	0.199783	0.87104	0.262966	0.00509177	0.0976558	0.0313039

5.6.3 Discussion

1. The high rate of arrivals (λ) generates a large number of customers in the system ($E(L) \nearrow$), which generates a high probability of busy period; the system remains active for a long period ($P_b \nearrow$), this leads to a loss of customers due to balking ($B_r \nearrow$), in return, the probability of working vacation for both type-1 and type-2 decreases (($P_{wv1}, P_{wv2}) \searrow$) which explains the reduction in the number of customers lost due to reneging ($R_r \nearrow$) when $\lambda > 0.8$, (high number of arrivals) (see Table 5.1 and Figure 5.4).

Table 5.2: Effect of ξ_1 on performance measures for $\lambda = 1.5$, $\eta = 0.6$, $\gamma_1 = 0.5$, $\gamma_2 = 0.7$, $\beta = 0.5$, $\theta_1 = 0.6$, $\theta_2 = 0.4$, $\mu = 4.0$, $\nu_1 = 3.5$, $\nu_2 = 2.5$, $\xi_2 = 0.3$, $c = 5$

ξ_1	E[L]	P_{id}	P_b	B_r	R_r	P_{wv1}	P_{wv2}
0.1	16.5815	0.177576	0.904246	0.0237715	0.00156507	0.0749637	0.0207898
0.4	16.5868	0.177145	0.903775	0.0951205	0.00384072	0.0752516	0.0209734
0.7	16.5893	0.176925	0.903524	0.166491	0.00589777	0.0753659	0.0211103
1.	16.5907	0.176781	0.903354	0.237872	0.00776433	0.0754207	0.0212249
1.3	16.5917	0.176674	0.903226	0.309259	0.00946517	0.0754489	0.0213249
1.6	16.5925	0.17659	0.903122	0.380651	0.0110212	0.0754633	0.0214142

Table 5.3: Effect of ξ_2 on performance measures for $\lambda = 1.5$, $\eta = 0.6$, $\gamma_1 = 0.5$, $\gamma_2 = 0.7$, $\beta = 0.5$, $\theta_1 = 0.6$, $\theta_2 = 0.4$, $\mu = 4.0$, $\nu_1 = 3.5$, $\nu_2 = 2.5$, $\xi_1 = 0.5$, $c = 5$

ξ_2	E[L]	P_{id}	P_b	B_r	R_r	P_{wv1}	P_{wv2}
0.1	16.5866	0.177117	0.903748	0.0237799	0.00405103	0.0753251	0.0209265
0.4	16.5881	0.177044	0.903657	0.0951295	0.0047841	0.0752939	0.0210494
0.7	16.5886	0.177022	0.903618	0.166484	0.00543695	0.0752845	0.0210977
1.	16.5889	0.177015	0.903595	0.23784	0.00602132	0.0752814	0.0211234
1.3	16.589	0.177013	0.90358	0.309196	0.00654725	0.0752807	0.0211394
1.6	16.589	0.177014	0.903568	0.380554	0.00702304	0.0752812	0.0211504

Table 5.4: Effect of θ_1 on performance measures for $\lambda = 1.5$, $\eta = 0.6$, $\gamma_1 = 0.5$, $\gamma_2 = 0.7$, $\beta = 0.5$, $\theta_2 = 0.4$, $\mu = 4.0$, $\nu_1 = 3.5$, $\nu_2 = 2.5$, $\xi_1 = 0.5$, $\xi_2 = 0.3$, $c = 5$

θ_1	E[L]	P_{id}	P_b	B_r	R_r	ECS
			$c = 1$			
0.1	24.4782	0.0292202	0.983983	1.29045	0.000778342	3.82455
0.3	24.3488	0.0325247	0.982171	0.998824	0.000866366	3.8047
0.5	24.1223	0.0375561	0.979413	0.708351	0.00100039	3.77449
0.7	23.683	0.0461467	0.974704	0.420257	0.00122922	3.72291
0.9	22.6544	0.0641399	0.964841	0.138449	0.0017085	3.61487
			$c = 5$			
0.1	19.6249	0.12074	0.934316	0.953115	0.00310224	14.9426
0.3	18.7041	0.137879	0.924992	0.705768	0.00354259	14.2843
0.5	17.4285	0.161521	0.91213	0.469016	0.00415004	13.3752
0.7	15.5485	0.196235	0.893245	0.250419	0.00504195	12.0387
0.9	12.5094	0.252175	0.862813	0.0667967	0.00647926	9.88298

2. Increasing impatience rates (ξ_1) and (ξ_2) leads the system to remain more in the idle state ((P_{wv1}, P_{wv2}) $\nearrow$), this can generate customer accumulation which results in an increase in ((B_r, R_r) $\nearrow$) (see Tables 5.2-5.3). This implies a decrease in the total revenue and consequently in the total expected profit (see Table 5.8).

Table 5.5: Effect of θ_2 on performance measures for $\lambda = 1.5$, $\eta = 0.6$, $\gamma_1 = 0.5$, $\gamma_2 = 0.7$, $\beta = 0.5$, $\theta_1 = 0.6$, $\nu_1 = 3.5$, $\nu_2 = 2.5$, $\xi_1 = 0.5$, $\xi_2 = 0.3$, $c = 5$

θ_2	E[L]	P_{id}	P_b	B_r	R_r	ECS
			$\mu = 4.0$			
0.1	17.2236	0.176859	0.929029	0.370651	0.000840898	13.236
0.3	16.8461	0.176883	0.913833	0.362384	0.00305724	12.9634
0.5	16.26	0.177411	0.890981	0.349546	0.00642186	12.5408
0.7	15.2587	0.179066	0.853054	0.327606	0.0120388	11.8195
0.9	13.2464	0.183805	0.778864	0.283498	0.0230487	10.3707
			$\mu = 6.0$			
0.1	14.7726	0.248736	0.900186	0.31786	0.00118264	17.06
0.3	14.3243	0.246891	0.879729	0.30803	0.00426725	16.5644
0.5	13.6418	0.244814	0.849561	0.293062	0.00886169	15.8108
0.7	12.511	0.242473	0.801021	0.268253	0.0163017	14.5632
0.9	10.3604	0.239995	0.711261	0.221058	0.0300949	12.1917

Table 5.6: Effect of η on performance measures for $\lambda = 1.5$, $\gamma_1 = 0.5$, $\gamma_2 = 0.7$, $\beta = 0.5$, $\theta_1 = 0.6$, $\theta_2 = 0.4$, $\mu = 4.0$, $\nu_1 = 3.5$, $\nu_2 = 2.5$, $\xi_1 = 0.5$, $\xi_2 = 0.3$, $c = 5$

η	E[L]	P_{id}	P_b	B_r	R_r	P_{wv1}	P_{wv2}
0.2	16.1663	0.236851	0.95705	0.0463332	0.00202851	0.0335763	0.00937388
0.5	16.5037	0.188986	0.914324	0.118295	0.00404643	0.0669772	0.0186988
0.8	16.7277	0.157215	0.885964	0.191888	0.00538585	0.0891477	0.0248883
1.1	16.8872	0.134588	0.865767	0.266407	0.00633974	0.104937	0.0292963
1.4	17.0065	0.117655	0.850652	0.341503	0.0070536	0.116753	0.0325951

Table 5.7: Effect of β on performance measures for $\lambda = 1.5$, $\eta = 0.6$, $\gamma_1 = 0.5$, $\gamma_2 = 0.7$, $\theta_1 = 0.6$, $\theta_2 = 0.4$, $\mu = 4.0$, $\nu_1 = 3.5$, $\nu_2 = 2.5$, $\xi_1 = 0.5$, $\xi_2 = 0.3$, $c = 5$

β	E[L]	P_{id}	P_b	B_r	R_r	P_{wv1}	P_{wv2}
0.1	16.5836	0.177482	0.904196	0.0237758	0.00445095	0.0749866	0.0208172
0.3	16.5857	0.177273	0.90394	0.0713363	0.0044995	0.0751417	0.0209182
0.5	16.5878	0.177059	0.903677	0.118909	0.00454926	0.0753002	0.0210224
0.7	16.5899	0.176839	0.903408	0.166495	0.00460028	0.075462	0.0211299
0.9	16.5921	0.176614	0.903132	0.214094	0.0046526	0.0756274	0.0212408

3. High service rate whether during the busy period or type-1/type 2 working vacation periods generates a significant average number of customers served (see Table 5.5 and Figure 5.3). Then, this (with the increases of λ) implies an increase in the total expected cost, along with the total revenue and total expected profit (see Table 5.8 and Figure

Table 5.8: Effect of c on total cost and total revenue for $\lambda = 1.5$, $\eta = 0.6$, $\gamma_1 = 0.5$, $\gamma_2 = 0.7$, $\beta = 0.5$, $\theta_1 = 0.6$, $\theta_2 = 0.4$, $\mu = 4.0$, $\nu_1 = 3.5$, $\nu_2 = 2.5$, $\xi_1 = 0.5$, $\xi_2 = 0.3$, $C_l = 20.0$, $C_i = 20.0$, $C_w = 16.0$, $C_r = 5.0$, $C_\mu = 15.0$, $C_{\nu_i} = 8.0$, Revenue = 100

	$c=2$		$c=5$		$c=8$		$c=11$	
	TCF	T_r	TCF	T_r	TCF	T_r	TCF	T_r
$\lambda = 1.0$	625.612	625.177	838.09	1094.42	1099.98	1325.74	1384.33	1448.12
$\lambda = 1.5$	661.12	679.007	886.945	1277.71	1149.88	1615.52	1432.03	1815.46
$\lambda = 2.0$	682.12	709.947	919.018	1397.51	1185.03	1819.65	1467.3	2087.64
$\mu = 4.0$	661.12	679.007	886.945	1277.71	1149.88	1615.52	1432.03	1815.46
$\mu = 4.5$	665.987	746.743	909.829	1375.32	1194.54	1716.05	1499.62	1911.17
$\mu = 5.0$	671.345	811.425	933.948	1464.78	1240.85	1805.94	1569.04	1995.35
$\nu_1 = 3.0$	653.11	678.912	866.935	1277.46	1117.87	1615.15	1388.03	1814.99
$\nu_1 = 3.5$	661.12	679.007	886.945	1277.71	1149.88	1615.52	1432.03	1815.46
$\nu_1 = 4.0$	669.13	679.087	906.955	1277.92	1181.88	1615.84	1476.03	1815.86
$\nu_2 = 2.0$	653.119	678.975	866.946	1277.64	1117.88	1615.41	1388.03	1815.33
$\nu_2 = 2.5$	661.12	679.007	886.945	1277.71	1149.88	1615.52	1432.03	1815.46
$\nu_2 = 3.0$	669.122	679.034	906.945	1277.78	1181.87	1615.61	1476.03	1815.57
$\beta' = 0.2$	661.142	679.033	887.003	1277.93	1149.95	1615.99	1432.11	1816.18
$\beta' = 0.5$	661.12	679.007	886.945	1277.71	1149.88	1615.52	1432.03	1815.46
$\beta' = 0.7$	661.106	678.991	886.908	1277.57	1149.83	1615.22	1431.98	1815.0
$\eta = 0.2$	656.113	672.156	879.596	1248.1	1142.36	1565.35	1424.96	1749.5
$\eta = 0.5$	660.133	677.656	885.479	1271.81	1148.37	1605.44	1430.6	1802.15
$\eta = 0.8$	662.752	681.24	889.385	1287.54	1152.4	1632.36	1434.42	1837.78
$\theta'_1 = 0.2$	631.196	641.009	839.621	1110.4	1100.55	1337.65	1384.66	1456.31
$\theta'_1 = 0.5$	670.576	691.202	903.83	1337.52	1168.7	1721.61	1450.92	1958.77
$\theta'_1 = 0.7$	683.806	708.539	929.451	1428.43	1198.62	1890.42	1481.97	2194.36
$\theta'_2 = 0.2$	631.038	639.497	843.002	1123.82	1104.68	1365.85	1388.79	1495.93
$\theta'_2 = 0.5$	656.841	673.375	880.197	1254.08	1142.63	1575.5	1424.89	1762.8
$\theta'_2 = 0.7$	664.419	683.356	892.265	1296.34	1155.67	1647.49	1437.79	1857.92
$\xi_1 = 0.2$	661.051	679.034	886.862	1277.59	1149.8	1615.29	1431.97	1815.18
$\xi_1 = 0.5$	661.12	679.007	886.945	1277.71	1149.88	1615.52	1432.03	1815.46
$\xi_1 = 0.8$	661.151	678.963	886.992	1277.73	1149.92	1615.61	1432.07	1815.6
$\xi_2 = 0.2$	661.111	679.009	886.935	1277.7	1149.87	1615.49	1432.02	1815.43
$\xi_2 = 0.5$	661.131	679.	886.958	1277.72	1149.89	1615.54	1432.04	1815.49
$\xi_2 = 0.8$	661.138	678.985	886.968	1277.72	1149.9	1615.55	1432.05	1815.51
$\gamma_1 = 0.2$	661.702	679.791	887.79	1281.12	1150.74	1621.3	1432.84	1823.07
$\gamma_1 = 0.4$	661.281	679.234	887.182	1278.68	1150.12	1617.17	1432.26	1817.63
$\gamma_1 = 0.6$	660.961	678.81	886.719	1276.83	1149.65	1614.02	1431.82	1813.5
$\gamma_2 = 0.8$	661.039	678.913	886.829	1277.28	1149.76	1614.78	1431.92	1814.49
$\gamma_2 = 1.2$	660.816	678.616	886.5	1275.97	1149.42	1612.54	1431.6	1811.53
$\gamma_2 = 1.5$	660.694	678.455	886.32	1275.26	1149.23	1611.33	1431.42	1809.93

5.5). As expected, we can further notice that large service rates during vacation periods ν_1 and ν_2 have the same impact as mu on TCF and T_r $((TCF, T_r) \nearrow)$, this has a nice impact on the total expected profit.

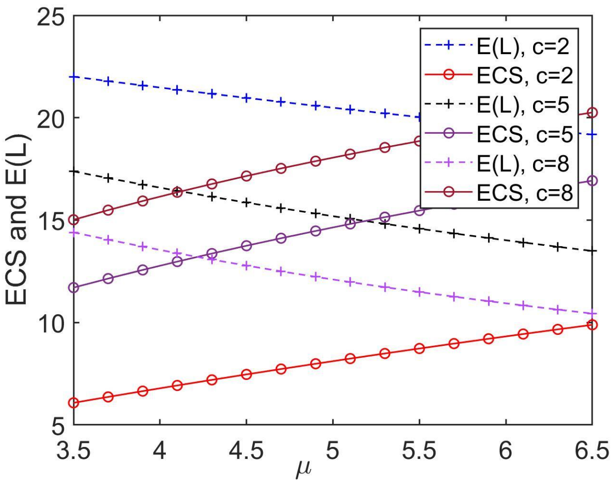

Figure 5.3: Variations in ECS and $E(L)$ with respect to μ and number of servers c.

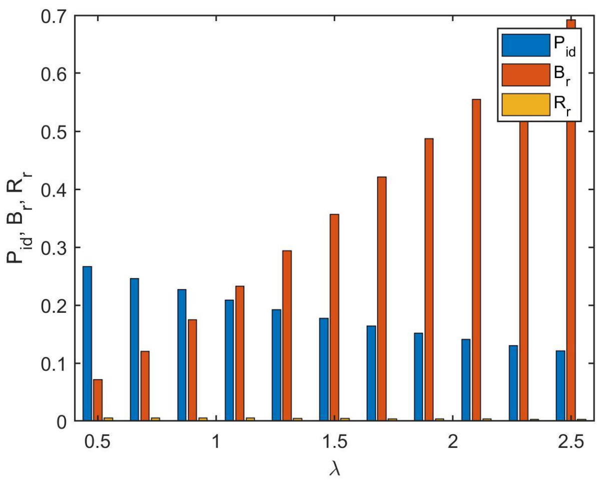

Figure 5.4: Impact of λ on P_{id}, B_r and R_r.

The same for the vacation rates γ_1 and γ_2 (see Table 5.8). This is well explained, the higher the vacation rates, the larger the mean number of customers served. This generates a significant total revenue and therefore an important total expected profit.

4. From Tables 5.4-5.5 and Figure 5.3, we notice that the increases in the number of servers (c) and the service rate during regular busy state μ has the same impact on the different system performance. With both c and μ; the mean number of customers in the system becomes

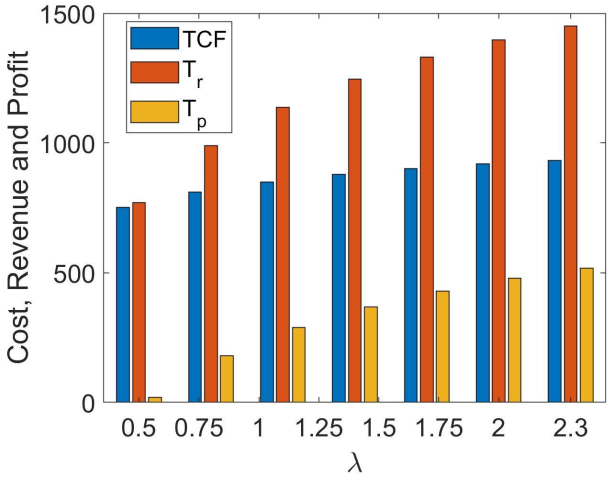

Figure 5.5: Impact of λ on TCF, T_r and T_p.

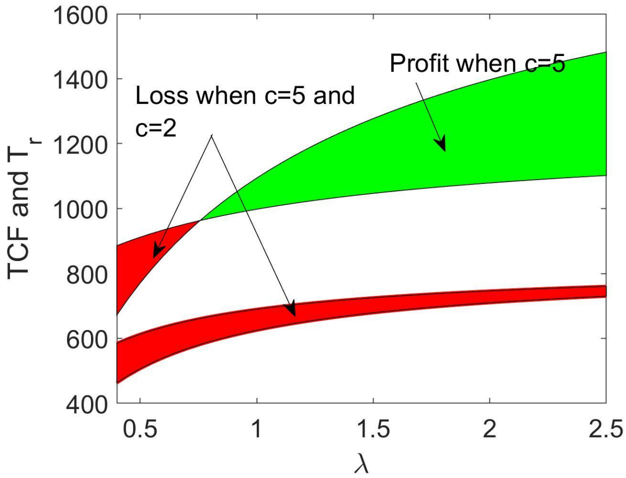

Figure 5.6: Impact of λ on TCF, T_r for different c.

less, which leads to a decrease in $(E(L))$ and therefore B_r decreases accordingly. In addition, P_b decreases and P_{id} grows which in turn implies an increase in (P_{wv1}, P_{wv2}) at which customers may renege to the impatience phenomenon $(R_r \nearrow)$.

5. Unlike joining probabilities θ_1 and θ_2, that have a negative impact on the mean number of customers served $(ESC \nearrow)$) (Tables 5.4-5.5), balking probabilities $\theta_1^{'}$ and $\theta_2^{'}$, have a nice impact on total revenue (see Table 5.8). This could be because of the chosen parameters and costs. Note that intentionally, for the selected parameters, we chose such cost

parameter values, to highlight the importance of the multi-server case.

6. Obviously, the increase in the number of servers in the system implies an increase in the total expected cost per unit time (see Table 5.8). On the other hand, the large number of c generates a significant increase in the mean number of customers served per unit time (ECS$\nearrow$) (see Figure 5.3 and Table 5.4) which generates a significant augmentation in the total revenue ($T_r \nearrow$) and consequently ($T_p \nearrow$) (Figure 5.6). However, as expected, it is clearly visible that if the system composes of two servers ($c = 2$), or the number of arrivals is ($\lambda = 1.0$), the system experiences a real loss of profit (see Table 5.8 and Figure 5.6).

7. It is well observed that waiting server rate η and non-interruption probability β have same impact on the different performance measures (see Tables 5.6-5.7):

    - Clearly, when the mean waiting server $(1/\eta)$ is small, the system switches quickly to the vacation state $((P_{wv1}, P_{wv2}) \nearrow)$ at which the customers could be served at slower rate, at this situation the number of customers in the system significantly accumulated $(E(L) \nearrow)$, this leads to high average balking and reneging rates.
    - Then, the increase of β means that the system decides to do not interrupt the working vacation process, this generates high working vacation probability for both type-1 and type-2 vacations that results in increasing the mean system size which in turn engenders a large loss $((B_r, R_r) \nearrow)$.
    - Therefore, increasing β', (resp. η) leads to a decrease in TCF and T_r (resp. an increase TCF and T_r) (see Table 5.8).

5.7 Conclusion

This investigation discussed a multi-server Markovian queue with the presence of differentiated working vacations, waiting servers, Bernoulli interruption schedule, and impatient customers 'balking and reneging'. The stability condition as well as the steady-state probabilities of the system were determined using the matrix geometric method. Various useful system performance measures were derived. The analytical results were approved numerically and graphically via Mathematica software.

The current research study can be extended in different directions by integrating batch arrival, Bernoulli feedback, retention of reneged customers, and many others.

Conclusion and future directions

Call centers play a pivotal role as essential service providers across various economic sectors, such as banking, healthcare, telecommunications, and more. However, managing call center operations is is a challenging task due to the high variability in customer arrivals, service times, and customer behavior. Queueing models emerge as powerful analytical tools for evaluating and optimizing call center performance, as they can capture the stochastic nature of customer and server dynamics, providing valuable insights for system management and congestion control. Nonetheless, most existing queueing models for call centers are based on simplifying assumptions that fail to reflect the intricate realities of call center operations, necessitating the development of more realistic and flexible models that can account for various factors influencing call center performance.

The primary objective of this thesis was to introduce novel queueing models that encapsulate the complex dynamics of call center operations, offering valuable insights for system management and congestion control. Specifically, two multi-server Markovian queueing models with K-variant working vacations and differentiated working vacations (DWV), subjected to Bernoulli schedule vacation interruption, with waiting servers, and impatience (balking, reneging) have been rigorously analyzed.

The key findings and contributions of this thesis are as follows:

- Analytical models: The queueing models were developed using the matrix-analytic method, enabling the derivation of steady-state probabilities and performance measures. This method leveraged matrix operations and linear algebra, providing analytical solutions for complex queueing systems.

- The thesis incorporated cost models and optimization techniques to determine optimal system parameters that minimize costs. This practical approach aids decision-makers in economically regulating queueing systems. Furthermore, a comprehensive cost-revenue-profit analysis has been presented.

- Real-world applications: The models presented in this thesis have practical applications in various domains, particularly call centers, highlighting the versatility and relevance of the proposed queueing models.
- Numerical illustrations: Extensive numerical results were provided to demonstrate the impact of different parameters on system performance, offering valuable insights for system managers to optimize operations and alleviate congestion.

The results demonstrated that the working vacation policy and its variants can significantly improve the performance and profitability of call center systems by allowing servers to perform secondary tasks at reduced service rates when the system is empty. Moreover, the results proved that system performance and the optimal solution are highly sensitive to various parameters, such as service rates, vacation rates, vacation interruption probability, balking rates, and reneging rates. Therefore, it is imperative for system managers and decision-makers to carefully monitor and adjust these parameters according to demand patterns, customer preferences, and operational costs.

Despite these notable contributions, there are certain limitations and open challenges that warrant further investigation:

- Complexity: Queueing models with working vacations and impatience can become increasingly complex, especially when considering multiple classes of customers and servers, heterogeneous service rates, and different types of working vacations. Consequently, finding analytical solutions for these models can be challenging and computationally intensive. Moreover, validating and calibrating these models with real-world data can be difficult and time-consuming.
- Generalization: The queueing models proposed in this thesis are based on specific assumptions and parameters that may not be applicable or realistic for some settings. For instance, the arrival process is assumed to follow a Poisson distribution, the service times are assumed to be exponentially distributed, and the working vacation rates are assumed to be constant. These assumptions may limit the generalization and robustness of the models and the results.
- Extensions: There are many possible extensions and directions for future research based on this thesis. Some of them are:
  - Incorporating customer feedback and satisfaction into the models, as they are important indicators of call center performance and quality.
  - Exploring other types of working vacation policies, such as adaptive working vacations, and comparing their effects on system performance and profitability.

– Presenting the transient analysis of the queueing systems studied in this thesis.

Bibliography

[1] Afanasyev, G. A. (2023). Cost optimization of queueing systems with vacations, *Theory of Probability & Its Applications, 68*(2). 10.1137/S0040585X97T99143X.

[2] Aissani, A., Taleb, S., Kernane, T., Said, G., Hamadouche, D. (2014). An $M/G/1$ retrial queue with working vacation, *Advances in Intelligent Systems and Computing*, 240, 443-452.

[3] Alfa, A. S. (2002). Discrete time queues and matrix-analytic methods, *Top*, 10(2), 147–210.

[4] Altman, E., Yechiali, U. (2006). Analysis of customers' impatience in queues with server vacations, *Queueing systems*, 52(4), 261–279.

[5] Ameur, L., Berdjoudj, L., Abbas, K. (2019). Sensitivity analysis of the M/M/1 retrial queue with working vacations and vacation interruption, *International Journal of Management Science and Engineering Management*, 14(4), 293-303.

[6] Ammar, S. I. (2017). Transient solution of an $M/M/1$ vacation queue with a waiting server and impatient customers, *Journal of the Egyptian Mathematical Society*, 25(3), 337–342.

[7] Anisimov, V., (2008). Switching processes in queueing models, *MGlaxoSmithKline, UK.*

[8] Arivudainambi, D., Godhandaraman, P., Rajadurai, P. (2014).Performance analysis of a single server retrial queue with working vacation, *OPSEARCH*, 51, 434-462.

[9] Asmussen, S. (1995). Stationary distributions for fluid flow models with or without Brownian noise, *Stoch. Model*, 11, 21–49.

[10] Asmussen, S. and Koole, G. (1993). Marked point processes as limits of Markovian arrival streams, *Journal of Applied Probability*, 30, 365-372.

[11] Ayyappan, G. and Nirmala, M. (2018). An $M[X]/G(a,b)/1$ queue with breakdown and delay time to two phase repair under multiple vacation, *Applications and Applied Mathematics*, 13(2), 639–663.

[12] Baba, Y. (2005). Analysis of a $GI/M/1$ queue with multiple working vacations, *Oper. Res. Lett.*, 33, 201-209.

[13] Baba, Y. (2012). The $M^X/M/1$ queue with multiple working vacation, *American Journal of Operations Research*, 2(2), 217-224.

[14] Baba, Y., (2010). The $M/PH/1$ queue with working vacation and vacation interruption, *Journal of System Sciences and Systems Engineering*, 19(4), 496-503.

[15] Badescu, A.L., Landriault, D. (2009). Applications of fluid flow matrix analytic methods in ruin theory–areview, *SerieA: Matemáticas de la Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales*, 103(2), 353–372.

[16] Bala Murugan, S. P., Keerthana, R. (2023). An $M/G/1$ Feedback retrial queue with working vacation and a waiting server, *Journal of Computational Analysis & Applications*, 31(1).

[17] Banik, A., (2010). Analysis of single server working vacation in $GI/M/1/N$ and $GI/M/1/\infty$ queueing systems, *International Journal of Operational Research*, 7(3), 314-333.

[18] Banik, A.D., Gupta, U.C., Pathak, S.S. (2007). On the $GI/M/1/N$ queue with multiple working vacations analytic analysis and computation, *Appl. Math. Model.*, 31, 1701-1710.

[19] Barbhuiya, F. P., Gupta, U. C. (2020). A discrete-time $GI^X/Geo/1$ queue with multiple working vacations under late and early arrival system, *Methodology and Computing in Applied Probability*, 22, 599-624.

[20] Begum, A., Choudhury, G. (2023). Analysis of an $M/G_1G_2/1$ queue with Bernoulli vacation and server breakdown, *International Journal of Applied and Computational Mathematics*, 9(1), 9.

[21] Bhat, U. N. and Basawa, I. V. (2002). Maximum likelihood estimation in queueing systems, *Advances on Methodological and Applied Aspects of Probability and Statistics*, (Ed. N. Balakrishnan), Taylor & Francis, N.Y., 13–29.

[22] Bhat, U.N. et al., (2011). A survey on working vacation queueing models, *International Journal Of Operational Research (IJOR)*, vol. 11(4), pp. 395-425 .

[23] Bini, D., Latouche, G., Meini, B. (2005). Numerical methods for structured Markov chains, *Oxford University Press on Demand*.

[24] Bouchentouf, A. A. and Guendouzi, A. (2020). The $M^X/M/c$ Bernoulli feedback queue with variant multiple working vacations and impatient customers: Performance and economic analysis, *Arabian Journal of Mathematics*, 9(2), 309–327.

[25] Bouchentouf, A. A., Boualem, M., Cherfaoui, M., Medjahri, L. (2021). Variant vacation queueing system with Bernoulli feedback, balking and server's states-dependent reneging, *Yugosl. J. Oper. Res*, 31, 557-575.

[26] Bouchentouf, A. A., Boualem, M., Yahiaoui, L., Ahmad, H., (2022). A multi-station unreliable machine model with working vacation policy and customers' impatience, *Quality Technology & Quantitative Management*, 19(6), 766-796.

[27] Bouchentouf, A. A., Cherfaoui, M., Boualem, M. (2021). Analysis and performance evaluation of Markovian feedback multi-server queueing model with vacation and impatience, *American Journal of Mathematical and Management Sciences*, 40(3), 261-282.

[28] Bouchentouf, A. A., Cherfaoui, M., Boualem, M. (2019). Performance and economic analysis of a single server feedback queueing model with vacation and impatient customers, *Opsearch*, 56(1), 300-323.

[29] Bouchentouf, A. A., Guendouzi, A. (2019). Cost optimization analysis for an $M^X/M/c$ vacation queueing system with waiting servers and impatient customers, *SeMA Journal*, 76, 309-341.

[30] Bouchentouf, A. A., Guendouzi, A. (2021). Single server batch arrival Bernoulli feedback queueing system with waiting server, K-variant vacations and impatient customers, *In Operations Research Forum*, Springer International Publishing, 2, 1-23.

[31] Bouchentouf, A. A., Guendouzi, A., Kandouci, A., (2018). Performance and economic analysis of Markovian Bernoulli feedback queueing system with vacations, waiting server and impatient customers, *Acta Univ. Sapientiae, Mathematica*, 10(2), 218-241.

[32] Bouchentouf, A. A., Guendouzi, A., Kandouci, A., (2019). Performance and economic study of heterogeneous $M/M/2/N$ feedback queue with working vacation and impatient customers, *ProbStat Forum*, 12(1), 15-35.

[33] Bouchentouf, A. A., Guendouzi, A., Majid, S. (2020). On impatience in $M/M/1/N/DWV$ queue with vacation interruption, *Croatian Operational Research Review*, 11(1), 21-37.

[34] Bouchentouf, A. A., Kadi, M. and Rabhi, A. (2014). Analysis of two heterogeneous server queueing model with balking, reneging and feedback, *Mathematical Sciences And Applications E-Notes*, 2(2), 10–21.

[35] Bouchentouf, A.A., Cherfaoui, M., Boualem, M., (2021). Analysis and performance evaluation of Markovian feedback multi-Server queueing model with vacation and impatience, *American Journal of Mathematical and Management Sciences*, 40(3), 261–282.

[36] Chakravarthy, S. R., (2018). Working vacation queueing models: theory and applications, *International Series In Operations Research & Management Science*, CRC Press LLC.

[37] Chakravarthy, S.R., (2023), Introduction to matrix-analytic methods in queues 1: analytical and simulation approach – basics, *International Series In Operations Research & Management Science*, Springer Nature Switzerland AG, 33-41.

[38] Chandrasekaran, V. M., Indhira, K., Saravanarajan, M. C., Rajadurai, P. (2016). A survey on working vacation queueing models, *International Journal of Pure and Applied Mathematics*, 106(6),

[39] Chen, F., Song, J.S. (2001). Optimal policies for multi-echelon inventory problems with Markov modulated demand, *Oper. Res.*, 49(2), 226–234.

[40] Chen, H.Y., Li. J.H., Tian, N.S. (2009). The $GI/M/1$ queue with phase-type working vacations and vacation interruption, *J Appl. Math. Comput.*, 30, 121-141.

[41] Choudhury, G.L., Lucantoni, D.M., (2023). Matrix-analytic methods – an algorithmic approach to stochastic models, John Wiley & Sons Ltd.

[42] Dai, J. G. (1998). Editorial introduction, *Queueing Systems*, 30(1-2), 1–3.

[43] Daigle, J. N. (2005). Queueing theory with applications to packet telecommunication, *New York: Springer Verlag.*

[44] Daley, D. J. (1965). General customer impatience in the queue $GI/G/1$, *Journal of Applied probability*, 2(1), 186–205.

[45] Danilyuk, E., Moiseeva, S. and Nazarov, A. (2019). Asymptotic analysis of retrial queueing system $M/GI/1$ with collisions and impatient calls, ITMM 2019: *Information Technologies and Mathematical Modelling. Queueing Theory and Applications*, 230–242.

[46] Deepa, B., Kalidass, K., (2018). The Markovian vacation queues with a waiting server and geometric abandonments, *International Journal of Pure and Applied Mathematics*, 118, 1903-1910.

[47] Do, T.V. (2010). $M/M/1$ retrial queue with working vacations, *Acta Informatica*, 47, 67-75.

[48] Doshi, B. T. (1986). Queueing systems with vacations-a survey, *Queueing Syst*,1(1), 29-66.

[49] Gao, S., Liu, Z. (2013). An $M/G/1$ queue with single working vacation and vacation interruption under Bernoulli schedule, *Applied Mathematical Modelling*, 37(3), 1564-1579.

[50] Gao, S., Wang, J., Li, W. (2014). An $M/G/1$ retrial queue with general retrial times, working vacations and vacation interruption, *Asia-Pacific Journal of Operational Research*, 31(2), 6-31.

[51] Gao, S., Yao, Y. (2014). An $M^X/G/1$ queue with randomized working vacations and at most J vacations, *International Journal of Computer Mathematics*, 91(3), 368-383.

[52] GnanaSekar, M. M. N., Kandaiyan, I., (2022). Analysis of an M/G/1 retrial queue with delayed repair and feedback under working vacation policy with impatient customers, *Symmetry*, 14(10), 2024.

[53] Goswami, V., Vijayalaxmi, P., Jyothsna, K. (2013). Analysis of $GI/M(n)/1/N$ queue with state-dependent multiple working, *OPSEARCH*, 50, 106-124.

[54] Goswami, V. (2014). A discrete-time queue with balking, reneging and working vacations, *Int. J. Stoch. Anal.*, Article ID 358529, 8 pages.

[55] Granville, K. (2019). An application of matrix analytic methods to queueing models with polling, University of Waterloo, Ontario, Canada.

[56] Shortle, J. F., Thompson, J. M., Gross, D., Harris, C. M. (2018). Fundamentals of queueing theory, *John Wiley & Sons*, Vol. 399.

[57] Guarguaglini, P., Marcoz, F. and Minguzzi, B. (1973). A Queuing Problem for an Intermittent Data Transmission System, *IEEE Transactions on Communications*, 21(3), 247–253.

[58] Gupta, G. K., Banerjee, A., Gupta, U. C. (2019). On finite-buffer batch-size-dependent bulk service queue with queue-length dependent vacation, *Quality Technology & Quantitative Management*, 17(5), 1-27.

[59] Gupta, P., Kumar, N.,(2021). Performance Analysis of Retrial Queueing Model with Working Vacation, Interruption, Waiting Server, Breakdown and Repair, *Journal of Scientific Research*, 13(3).

[60] Haghighi, A. M., (2008). Queueing models in industry and business, *Nova Science Publisher*.

[61] Haghighi, A.M., Mishev, D.P., (2014). Queueing models in industry & business, *Nova Science Pub Inc; 2nd edition.*

[62] Haghighi, A.M., Mishev, D.P., (2016). Delayed and network queues, *John Wiley and Sons Inc., New Jersey.*

[63] Haghighi, A.M., Mishev, D.P., (2016). Stochastic modeling in industry and management, *Modeling and Simulation in Industrial Engineering, Springer, Cham,* 131-178.

[64] Haight, F. A. (1957). Queueing with balking, *Biometrika*, 44(3/4), 360-369.

[65] Harrison, J. M. and Nguyen, V. (1990). The QNET method for two-moment analysis of open queueing networks, *Queueing Systems*, 6(1), 1–32.

[66] Harrison, J. M. and Nguyen, V. (1993). Brownian models of multiclass queueing networks: Current status and open problems, *Queueing Systems*, 13(1-3), 5–40.

[67] He, Q. M. (2014). Fundamentals of matrix-analytic methods, *Springer*, New York.

[68] He, Q.M. (1996). Queues with marked customers, A*dv. Appl. Probab*, 28, 567–587.

[69] He, X. and Hu, W. (2017). A Two-stage queue model to optimize layout of urban drainage system considering extreme rainstorms, *Mathematical Problems in Engineering*, Article ID 6380521, 18 pages, https://doi.org/10.1155/2017/6380521.

[70] Houalef, M., Bouchentouf, A.A., Yahiaoui, L., (2023). A Multi-server queue in a multi-phase random environment with waiting servers and customers' impatience under synchronous working vacation policy, *Journal of the Operations Research Society of China*, 11(3), 459-487

[71] Ibe, O. and Isijola, O. (2014). $M/M/1$ multiple vacation queueing systems with differentiated vacations, *Modelling and Simulation in Engineering*, Volume 2014, Article ID 158247, 6 pages, http://dx.doi.org/10.1155/2014/158247.

[72] Isijola, O. and Ibe, O. (2014). $M/M/1$ multiple vacation queueing systems with differentiated vacations and vacation interruptions, *IEEE Access*, 2, 1384–1395.

[73] Jain, M., Agrawal, P. K. (2007). $M/E_k/1$ queueing system with working vacation, *Quality Technology & Quantitative Management*, 4,455-470.

[74] Jain, M., Jain, A. (2010). Working vacations queueing model with multiple types of server breakdown, *Applied Mathematical Modelling*, 34(1), 1-13.

[75] Jain, M., Kumar, A. (2023). Unreliable server $M^{[X]}/G/1$ retrial feedback queue with balking, working vacation and vacation interruption, *Proceedings of the National Academy of Sciences*, India Section A: Physical Sciences, 93(1), 57-73.

[76] Jain, M., Upadhyaya, S., (2011). Synchronous working vacation policy for finite-buffer multiserver queueing system, *Applied Mathematics and Computation*, 217(24), 9912-9916.

[77] Jiang, T., Liu, L. (2017). The $GI/M/1$ queue in a multi-phase service environment with disasters and working breakdowns, *International Journal of Computer Mathematics*, 94(4), 707-726.

[78] Joshi, P. K., Gupta, S. and Rajeshwari, K. N. (2020). An study of steady state analysis of D/M/1 model and $M/G/1$ model with multiple vacation queueing systems, *South East Asian Journal of Mathematics and Mathematical Sciences*, 16(1), 37–50.

[79] Kao, E., Narayanan, K.S. (1991). Analyses of an $M/M/n/queue$ with server's vacations, *European Journal of Operational Research*, 54(2), 256-266.

[80] Kasim, E. (2016). Semigroup methods for the $M/G/1$ queueing model with working vacation and vacation interruption, *J. Math. Res*, 8(5), 56-77.

[81] Ke, J. (2007). Operating characteristic analysis on the $M^X/G/1$ system with a variant vacation policy and balking, *Applied Mathematical Modelling*, 31(7), 1321–1337.

[82] Ke, J. C., Wu, C. H., Zhang, Z. G., (2010). Recent developments in vacation queueing models: a short survey, *International Journal of Operations Research*, 7, 3-8.

[83] Ke, J.-C., Chang, F.-M., Liu, T.-H. (2019). $M/M/c$ balking retrial queue with vacation, *Quality Technology & Quantitative Management*, 16(1), 54-66.

[84] Keerthiga, S., Indhira, K. (2023). Two phase of service in $M/G/1$ queueing system with retrial customers, *The Journal of Analysis*, 1-27.

[85] Kelly, F. P. (1979). Reversibility and stochastic networks, *John Wiley and Sons, New York.*

[86] Kendall, D.G. (1953). Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain, *Annals of Mathematical Statistics*, 338-354.

[87] Kim, J. D. (2003). Analysis of queue-length distribution of the M/G/1 queue with working vacations ($M/G/1/WV$), *In Proceedings of Hawaii International Conference on Statistics and Related Fields*, 1191-1200.

[88] Kingman, J. F. C. (1969). Markov population processes, *Journal of Applied Probability*, 6(1), 1–18.

[89] Kleinrock, L. (1975). Queueing systems, *John Wiley, New York*, Vol. I: Theory.

[90] Kleinrock, L. (1976). Queueing systems, *John Wiley, New York*, Vol. II: Computer Applications.

[91] Kleinrock, L. (1979). Sorbanállás, kiszolgálás : Bevezetés a tömegkiszolgálási rendszerek elméletébe, *M'uszaki Kiadó.*

[92] Kroese, D.P., Scheinhardt, W.R.W., Taylor, P.G. (2004). Spectral properties of the tandem Jackson network,seenasaquasi-birth-and-deathprocess, *Ann.Appl.Probab.*. 14(4),2057–2089.

[93] Kumar, R. (2013). Economic analysis of an $M/M/c/N$ queuing model with balking, reneging and retention of reneged customers, *OPSEARCH*, 50(3), 383–403.

[94] Kumar, R. (2016). A single-server Markovian queuing system with discouraged arrivals and retention of reneged customers, *Yugoslav Journal of Operations Research*, 24(1), 119–126.

[95] Kumar, R. and Sharma, S. K. (2012). $M/M/1/N$ queuing system with retention of reneged customers, *Pakistan Journal of Statistics and Operation Research*, 8(4), 859–866.

[96] Kumar, R. and Sharma, S. K. (2012). An $M/M/1/N$ queuing model with retention of reneged customers and balking, *American Journal of Operational Research*, 2(1), 1–5.

[97] Kumar, R. and Sharma, S. K. (2013). An $M/M/c/N$ queuing system with reneging and retention of reneged customers, *International Journal of Operational Research*, 17(3), 333–334.

[98] Kumar, R. and Sharma, S. K. (2018). Transient analysis of an $M/M/c$ queuing system with balking and retention of reneging customers, *Communications in Statistics-Theory and Methods*, 47(6), 1318–1327.

[99] Kumar, R. and Sharma, S. K. (2018). Transient performance analysis of a single server queuing model with retention of reneging customers, *Yugoslav Journal of Operations Research*, 28(3), 315–331.

[100] Kumaran, S. K., Dogra, D. P. and Roy, P. P. (2019). Queuing theory guided intelligent traffic scheduling through video analysis using Dirichlet process mixture model, *Expert Systems with Applications*, 118(2019), 169–181.

[101] Latouche, G., & Ramaswami, V. (1999). Introduction to matrix analytic methods in stochastic modeling, *ASA-SIAM Series on Statistics and Applied Probability*, Philadelphia.

[102] Laxmi, P. V. and Jyothsna, K. (2013). Analysis of finite buffer renewal input queue with balking and multiple working vacations, *OPSEARCH*, 50(4), 548–565.

[103] Laxmi, P. V. and Kassahun, T. W. (2017). A multi-server infinite capacity Markovian feedback queue with balking, reneging and retention of reneged customers, *International Journal of Mathematical Archive*, 8(2), 53–59.

[104] Laxmi, P. V. and Rajesh, P. (2016). Analysis of variant working vacations on batch arrival queues, *OPSEARCH*, 53(2), 303–316.

[105] Laxmi, P. V. and Rajesh, P. (2017). Performance measures of variant working vacations on batch arrival queue with reneging, *International Journal of Mathematical Archive*, 8(8), 85–96.

[106] Laxmi, V. P., Rajesh, P. (2020). Performance characteristics of discrete-time queue with variant working vacations, *International Journal of Operations Research and Information Systems* (IJORIS), 11(2), 1-21.

[107] Lee, D.H., Kim, B.K., (2015). A note on the sojourn time distribution of an $M/G/1$ queue with a single working vacation and vacation interruption, *Operations Research Perspectives*, 2, 57-61.

[108] Levy, Y., Yechiali, U. (1975). Utilization of idle time in an $M/G/1$ queueing system, *Management Science*, 22(2), 202-211.

[109] Li, J. H., Tian, N. and Ma, Z. Y. (2008). Performance analysis of $GI/M/1$ queue with working vacations and vacation interruption, *Applied Mathematical Modelling*, 32: 2715-2730.

[110] Li, J., Tian, N. (2007). The $M/M/1$ queue with working vacations and vacation interruptions, *J Syst Sci Syst Eng*, 16, 121-127.

[111] Li, J., Tian, N. (2007). The discrete-time $GI/Geo/1$ queue with working vacations and vacation interruption, *Appl. Math. Comput.*, 185, 1-10.

[112] Li, J., Tian, N. (2011). Performance analysis of a $GI/M/1$ queue with single working vacation, *Appl. Math. Comput.*, 217, 4960-4971.

[113] Li, J., Tian, N., Ma, Z. (2008) . Performance analysis of $GI/M/1$ queue with working vacations and vacation interruption, *Appl. Math. Model.*, 32, 2715-2730.

[114] Li, P., Xu, X., Cao, X. (2012). Analysis for the $M/G/1$ working vacation queue with set-up period, *Journal of Information and Computational Science*, 9, 801-812.

[115] Lin, C.H., Ke, J. C., (2009). Multi-server system with single working vacation, *Applied Mathematical Modelling*, 33(7), 2967-2977.

[116] Liu, B., Cui, L., Wen, Y., Shen, J. (2015). A cold standby repairable system with working vacations and vacation interruption following Markovian arrival process, *Reliability Eng. Sys. Safety*, 142, 1-8.

[117] Liu, W. Y., Xu, X. L. and Tian, N. (2007). Stochastic decompositions in the $M/M/1$ queue with working vacations, *Operations Research Letters*, 35: 595-600.

[118] Liu, Y. , Ye, Q., Yan, J., (2023). Analysis of a $GI^X/M/1$ queue with two-stage vacation policy using shift operator method, *Communications in Statistics- Theory and Methods*, 1-30. DOI: 10.1080/03610926.2023.2250488.

[119] Lothar, B. and Dieter, B. (2005). An introduction to queueing theory and matrix-analytic methods, *Springer Verlag*, ISBN: 1402036302.

[120] Lucantoni, D.M., Choudhury, G.L., Whitt, W. (1994). The transient $BMAP/G/1$ queue, *Stoch. Model*, 10, 145–182.

[121] Madheswari, S. P., Suganthi, P. and Josephine, S. A. (2016). Retrial queueing system with retention of reneging customers, *International Journal of Pure and Applied Mathematics*, 106(5), 11–20.

[122] Majid, S. , Manoharan, P. , (2017). Analysis of the M/M/1 queue with single working vacation and vacation interruption, *International Journal of Mathematics Trends and Technology*, 47 (1), 32-40.

[123] Majid, S., Manoharan, P., (2018). Impatient customers in an $M/M/c$ queue with single and multiple synchronous working vacations, *Pakistan Journal of Statistics and Operation Research*, 14(3), 571-594.

[124] Majid, S., Manoharan, P., (2019). Analysis of an $M/M/1$ queue with working vacation and vacation interruption, *Applications and Applied Mathematics*, 14(1), 19-33.

[125] Meena, R. K., Jain, M., Assad, A., Sethi, R., Garg, D. (2022). Performance and cost comparative analysis for $M/G/1$ repairable machining system with N-policy vacation, *Mathematics and Computers in Simulation*, 200, 315-328.

[126] Melikov, A. Z., Shahmaliyev, M. O., Nair, S. S., (2023). Matrix-geometric method for the analysis of a queuing system with perishable inventory, *International Journal of Operations Research and Information Systems*, 14(2), 1-18.

[127] Murugan, S. P. B., Keerthana, R. (2023). An $M/G/1$ Feedback retrial queue with working vacation and a waiting server, J*ournal of Computational Analysis and Applications*, 61.

[128] Nandy, N., & Pradhan, S. (2021). On the joint distribution of an infinite-buffer discrete-time batch-size-dependent service queue with single and multiple vacations. *Quality Technology & Quantitative Management*, 18(4), 432-467.

[129] Nandy, N., Pradhan, S. (2023). Stationary joint distribution of a discrete-time group-arrival and batch-size-dependent service queue with single and multiple vacation, *Communications in Statistics–Theory and Methods*, 52(9), 3012-3046.

[130] Naoumov, V., Krieger, U.R., Wagner, D. (1997). Analysis of a multiserver delay-loss system with a general Markovian arrival process, *Matrix-Analytic Methods in Stochastic Models.*

[131] Neuts, M. F. (1981). Matrix-geometric solutions in stochastic models: an algorithmic approach, *The Johns Hopkins University Press*, Baltimore.

[132] Neuts, M. F. (1989). Structured stochastic matrices of $M/G/1$ type and their applications, *Marcel Dekker*, New York.

[133] Bean, N. G., O'Reilly, M. M., Palmowski, Z. (2022). Matrix-analytic methods for the analysis of stochastic fluid-fluid models, *Stochastic Models*, 38(3), 416-461.

[134] Padmavathy, R., Kalidass, K. and Ramanath, K. (2011). Vacation queues with impatient customers and a waiting server, *Int. J. Latest Trends Softw. Eng*, 1(1), 10–19.

[135] Plateau, B., & Atif, K. (1991). Stochastic automata network for modeling parallel systems, *IEEE Transactions on Software Engineering*, 17(10), 1093-1108.

[136] Rajadurai, P. (2018). Sensitivity analysis of an $M/G/1$ retrial queueing system with disaster under working vacations and working breakdowns, *RAIRO-Operations Research*, 52(1), 35-54.

[137] Rajadurai, P., Saravanarajan, M. C., Chandrasekaran, V. M. (2018). A study on $M/G/1$ feedback retrial queue with subject to server breakdown and repair under multiple working vacation policy, *Alexandria Engineering Journal*, 57(2), 947-962.

[138] Rajadurai, P., Saravanarajan, M.C., Chandrasekaran, V.M. (2015). Analysis of an $M/G/1$ retrial queue with balking, negative customers, working vacations and server breakdown, I*nt. J. Appl. Eng. Res.*, 10, 4130-4135.

[139] Ross, S.M. (2014). Introduction to probability models, *Academic press*, San Diego.

[140] Schassberger, R.S. (1970). On the waiting time in the queueing system $GI/G/1$, *Ann. Math. Statist.*, 41, 182-187.

[141] Schassberger, R.S. (1973). Warteschlangen, *Springer-Verlag*, Berlin.

[142] Seeniraj, G., Moganraj, D., Jayaraman, R. (2023, June). A steady state analysis of single server $M/G/1$ queueing system with multiple working vacations, *In AIP Conference Proceedings* (Vol. 2649, No. 1). AIP Publishing.

[143] Seenivasan, M., Abinaya, R. (2022). Markovian queueing model with single working vacation and catastrophic, *Materials Today: Proceedings*, 51, 2348-2354.

[144] Selvaraju, N., Goswami, C. (2013). Impatient customers in an $M/M/1$ queue with single and multiple working vacations,*Computers & Industrial Engineering*, 65(2), 207-215.

[145] Servi, L. D., Finn, S.G. (2002). $M/M/1$ queues with working vacations $M/M/1/WV$, *Perform. Eval*, 50, 41-52.

[146] Shortle, John F. and Thompson, James M. and Gross, Donald and Harris, Carl M. , (2018). Fundamentals of queueing theory, 5th edition, Wiley, Hoboken, NJ.

[147] Som, B. K. and Seth, S. (2017). $M/M/c/N$ queueing systems with encouraged arrivals, reneging, retention and feedback customers, *Advanced Modeling and Optimization*, 19(1), 97–104.

[148] Sreenivasan, C., Srinivas, R., Chakravarthy, A., Krishnamoorthy, A. (2013). $MAP/PH/1$ queue with working vacations, vacation interruptions and N policy, *Applied Mathematical Modelling*, 37, 3879-3893.

[149] Srinivas, S. S. and Marathe, R, R. (2020). Equilibrium in a finite capacity $M/M/1$ queue with unknown service rates consisting of strategic and non-strategic customers, *Queueing Systems*, DOI: https://doi.org/10.1007/s11134-020-09671-x.

[150] Sudesh, R. and Azhagappan, A. (2020). Transient solution of an $M/M/\infty$ queue with system's additional tasks and impatient customers, *International Journal of Mathematics in Operational Research*, 16(1), 82–97.

[151] Sudhesh, R., Shapique, A., Dharmaraja, S. (2022). Analysis of a multiple dual-stage vacation queueing system with disaster and repairable server. *Methodology and Computing in Applied Probability.*

[152] Suranga Sampath, M., Liu, J., (2020). Impact of customers' impatience on an $M/M/1$ queueing system subject to differentiated vacations with a waiting server, *Quality Technology & Quantitative Management*, 17(2), 125–148.

[153] Tadj, L., Choudhury, G. (2005). Optimal design and control of queues, *Top,* 13(2), 359-412.

[154] Takagi, H. (1991). Queueing analysis: A foundation of performance evaluation (Vol. 1). *North-Holland.*

[155] Tao, L., Liu, Z., Wang, Z. (2011). The $GI/M/1$ queue with start-up period and single working vacation and Bernoulli vacation interruption, *Appl. Math. Comput.*, 218, 4401-4413.

[156] Tao, L., Wang, Z., Liu, Z. (2013). The GI/M/1 queue with Bernoulli-schedule-controlled vacation and vacation interruption. Applied Mathematical Modelling, 37(6), 3724-3735.

[157] Thakur, S., Jain, A., Ahuja, A., (2023). Analysis of MAP/PH/1 model with working vacation, working breakdown and two-phase repair. *Arab J Sci Eng.* https://doi.org/10.1007/s13369-023-08237-2.

[158] Tian, N., Li, J., Zhang, Z.G. (2009). Matrix analytic method and working vacation queues a survey, *Int. Journal of Information and Management Sciences*, 20, 603-633.

[159] Tian, N., Zhang, Z. G. (2006). Vacation queueing models: theory and applications, *Springer Science* & *Business Media.*

[160] Tian, R. Hu, L., Wu, X., (2016). Equilibrium and optimal strategies in $M/M/1$ queues with working vacations and vacation interruptions, *Mathematical Problems in Engineering*, 2016 Article ID 9746962, 10 pages.

[161] Tian, R., Zhang, Z. G., Su, S. (2021). On Markovian queues with single working vacation and Bernoulli interruptions, *Probability in the Engineering and Informational Sciences*, 1-28.

[162] Vadivukarasi, M., Kalidass, K., Jayaraman, R. (2022). Discussion on the optimization of finite buffer Markovian queue with differentiated vacations, *Soft Computing: Theories and Applications*, Springer, Singapore, 523-534.

[163] Van Vuuren, M. (2007). Performance analysis of manufacturing systems: queueing approximations and algorithms.

[164] Vignesh, P., Srinivasan, S., Sundari, S. M., Swedheetha, C. (2023). Study on the queuing characteristics of $M^X/G/1$ queuing model of stages of service with multi vacation policy and setup time phase, *International Journal of Process Management and Benchmarking*, 15(1), 25-45.

[165] Vijaya Laxmi, P. Goswami, V., Suchitra, V. (2013). Analysis of $GI/M(n)/1/N$ queue with single working vacation and vacation interruption, *Int. J. Math., Comput., Phy., Elec., Comp. Eng.*, 7, 740-746.

[166] Vijaya Laxmi, P., Goswami, V., Jyothsna, K. (2013). Analysis of finite buffer Markovian queue with balking, reneging and working vacations, *International Journal of Strategic Decision Sciences* (IJSDS), 4(1), 1-24.

[167] Vijaya Laxmi, P., Jyothsna, K. (2014). Performance analysis of variant working vacation queue with balking and reneging, *Int. J. of Math. Oper. Res.*, 6.

[168] Vijaya Laxmi, P., Kassahun, T. W., Bhavani, E. G. (2019, October). Analysis of a Markovian queueing system with single working vacation and impatience of customers, *Journal of Physics: Conf. Series*, 1344, 1-9.

[169] Vijaya Laxmi, P., Rajesh, P., Kassahun, T. W. (2019). Performance measures of variant working vacations on batch arrival queue with server breakdowns, *International Journal of Management Science and Engineering Management*, 14(1), 53-63.

[170] Vijaya Laxmi, P., Rajesh, P., Kassahun, T. W., (2021). Analysis of a variant working vacation queue with customer impatience and server

breakdowns, *International Journal of Operational Research*, 40 (4), 437-459.

[171] Vijaya Laxmi, P., Suchitra, V. (2014). Analysis of general input state dependent working vacation queue with changeover Time, *ISRN Computational Mathematics*, 2014, 1-8.

[172] Vuuren, van, M. (2007). Performance analysis of manufacturing systems: queueing approximations and algorithms, Technische Universiteit Eindhoven. https://doi.org/10.6100/IR625074

[173] Wang, J., Cui, S. and Wang, Z. (2018). Equilibrium strategies in $M/M/1$ priority queues with balking, *Production and Operations Management*, 28(1), 43–62.

[174] Wang, S., Xu, X. (2021). Equilibrium strategies of the fluid queue with working vacation, *Operational Research*, 21(2), 1211-1228.

[175] Whitt, W. (1983a). The queueing network analyzer, *The Bell System Technical Journal*, 62 (9), 2779–2815.

[176] Whitt, W. (1983b). Performance of the queueing network analyzer, *The Bell System Technical Journal*, 62 (9), 2817–2843.

[177] Whittle, P. (1967). Nonlinear migration processes, *Bull. Int. Stat. Inst.*, 42, 642–646.

[178] Whittle, P. (1968). Equilibrium distributions for an open migration process, *Journal of Applied Probability*, 5 (3), 567–571.

[179] Wolff, R.W., (2020). Stochastic modeling and the theory Of queues, (2nd ed.), Prentice Hall Inc.

[180] Wu, D., Takagi, H. (2006). $M/G/1$ queue with multiple working vacations, *Perform. Eval*, 63, 654-681.

[181] Xia, L., He, Q.M., Alfa, A.S. (2017). Optimal control of state-dependent service rates in a $MAP/M/1$ queue, *IEEE Trans. Autom. Control*. 62(10), 4965–4979.

[182] Xu, X. L., Liu, M. X. and Zhao, X. H. (2009). The bulk input $M^x/M/1$ queue with working vacations, *Journal of Systems Science and Systems Engineering*, 18(3) : 358-368.

[183] Xu, X. L., Zhang, Z. J., Tian, N. S. (2009). Analysis for the $M^X/M/1$ working vacation queue, *International journal of information and management sciences*, 20(3), 379-394.

[184] Yahiaoui, L., Bouchentouf, A. A., Kadi, M. (2019). Optimum cost analysis for an Geo/Geo/c/N feedback queue under synchronous working vacations and impatient customers. Croatian Operational Research Review, 211-226.

[185] Yang, D. Y. and Wu, Y. Y. (2017). Analysis of a finite-capacity system with working breakdowns and retention of impatient customers, *Journal of Manufacturing Systems*, 44(1), 207–216.

[186] Yang, D. Y., Wu, C. H. (2019). Performance analysis and optimization of a retrial queue with working vacations and starting failures, *Mathematical and Computer Modelling of Dynamical Systems*, 25(5), 463-481.

[187] Yang, D.Y. C.H. Wu. (2015). Cost-minimization analysis of a working vacation queue with N-policy and server breakdowns, *Computers & Industrial Engineering,* 82, 151-158.

[188] You, L., Wu, Q., Wei, J., Hu, J., Wang, J. and Liang, Y. (2020). A study of pedestrian evacuation model of impatient queueing with cellular automata, *Physica Scripta*, 95(9), 095211.

[189] Yue, D., Yue, W. (2009). Analysis of an $M/M/c/N$ queueing system with balking, reneging, and synchronous vacations, *Advances in queueing theory and network applications*, Springer, New York, NY, 165-180.

[190] Yue, D., Yue, W., Saffer, Z. and Chen, X. (2014). Analysis of an$M/M/1$ queueing system with impatient customers and a variant of multiple vacation policy, *Journal of Industrial and Management Optimization*, 10(1), 89–112.

[191] Yue, D., Yue, W., Xu, G., (2012). Analysis of customers impatient in an $M/M/1$ queue and working vacations, J*ournal of Industrial and Management Optimization*, 8(4), 895-908.

[192] Yue, W., Takahashi, Y. , Wittevrongel, S., Phung-Duc, T., (2019). Queueing theory and network applications, *in the 14th International Conference on Queueing Theory and Network Applications*, Springe.

[193] Zhang, H. (2018). An analysis of the $M^X/M/1$ queue with multiple working vacations by $GI/M/1$ type Markov process, *Acta Informatica*, 55(7), 613-624.

[194] Zhang, M., Hou, Z. (2010). Performance analysis of $M/G/1$ queue with working vacations and vacation interruption, *Journal of Computational and Applied Mathematics*, 234(10), 2977-2985.

[195] Zhang, M., Hou, Z. (2011). Performance analysis of $MAP/G/1$ queue with working vacations and vacation interruption, *Applied Mathematical Modelling*, 35, 1551-1560.

[196] Zhang, M., Hou, Z. (2012). $M/G/1$ queue with single working vacation, *Journal of Applied Mathematics and Computing* , 39, 221-234.

[197] Zhang, M., Hou, Z.(2011). Steady state analysis of $GI/M/1/N$ queue with a variant of multiple working vacations, *Comput. Ind. Eng.* 61, 1296-1301.

[198] Zhao, G., Cui, G. (2011). Approximation on $MAP/G/1$ queue with N-policy working vacation, *J. Conv. Inf. Tech.*, 6, 110-118.

[199] Zhao, G., Du, X., Tian, N. (2009). $GI/M/1$ Queue with Setup Period and Working Vacation and Vacation Interruption, *Int. J. Inform. Manage. Sci.*, 20, 351-363.

[200] Zukerman, M. (2013). Introduction to queueing theory and stochastic teletraffic models, *arXiv preprint arXiv:1307.2968.*

ملخص:

تسعى الرسالة إلى إدخال نماذج انتظار جديدة لمعالجة تحديات عمليات مراكز الاتصال. تم تحليل نموذجي انتظار ماركوفيين بنجاح مع تبني سياسة العمل خلال الإجازات، مما يسمح للخوادم بأداء مهام ثانوية بمعدلات خدمة منخفضة عندما يكون النظام فارغًا يعتبر هذا النهج مهمًا لتحسين أداء مراكز الاتصال وضمان تجربة عملاء سلسة وفعالة .

الأهداف الرئيسية لهذا البحث تشمل استخراج احتمالات الحالة الثابتة ومقاييس الأداء لهذه النظم باستخدام الأسلوب المصفوفي الهندسي. بالإضافة إلى ذلك، نقوم بتطوير نمادج التكلفة وإجراء تحليلات حساسية لتقييم تأثير مختلف معلمات النظام على مقاييس الأداء بشكل شامل.

النموذج الأول الذي تم فحصه هو نظام انتظار M/M/c الذي يشمل خوادم انتظار ورفض وتراجع، و إجازات نشطة متغيرة K، ويخضع لانقطاع إجازة جدول برنولي. بعد حساب الاحتمالات الثابتة للنظام، تم تطوير نموذج التكلفة باستخدام طريقة البحث المباشر لتحديد معدلات الخدمة المثلى خلال فترات العمل والإجازات بهدف تقليل التكاليف.

النموذج الثاني يدرس نظام انتظار M/M/c مع إجازات نشطة متباينة (DWV) ، واستخدام جدول برنولي لانقطاع الإجازات النشطة ونفاد صبر العملاء. يركز النموذج على إجازات طويلة (نوع 1) بعد خدمة عميل وإجازات قصيرة (نوع 2) بعد العودة. بعد الحالة المستقرة، يتم تحسين الأداء بواسطة دالة التكلفة وتحليل للتكاليف والإيرادات والأرباح، لضمان التوازن بين جودة الخدمة وكفاءة النظام وتكاليف التشغيل.

الكلمات المفتاحية: نماذج الطابور الماركوفية، إجازات نشطة، مقاييس الأداء، نفاد صبر العملاء، نموذج التكلفة.

Résumé :

La principale contribution de cette thèse est l’introduction de modèles de files d’attente markoviens innovants qui capturent la dynamique complexe des opérations de centres d'appels. Ces modèles utilisent la politique de vacances actives (working vacation), qui permet aux serveurs de faire des tâches secondaires à des taux de service réduits quand le système est inactif, une stratégie courante et efficace dans les centres d’appels, où les serveurs peuvent gérer différents types d’appels ou d’autres activités pendant les périodes d’inactivité.

Les objectifs principaux de cette recherche sont de dériver les probabilités à l’état stable et les mesures de performance pour ces systèmes d’attente, en utilisant la méthode de matrice géométrique. Nous développons aussi des modèles de coûts et faisons des analyses de sensibilité pour évaluer l’impact de divers paramètres du système sur les mesures de performance.

En premier lieu, nous considérons un système de files d'attente M/M/c avec vacances actives (working vacation), interruption de vacances selon une loi de Bernoulli, et clients impatients (balking et reneging). Ce système permet aux serveurs de prendre des vacances chaque fois que le système est vide après une période d'attente aléatoire (waiting servers), avec un nombre maximal de K vacances consécutives. Les probabilités à l'état stable du système sont obtenues, et la fonction de coût est optimisée par une méthode de recherche directe pour trouver les taux de service optimaux pendant les périodes de vacances actives et de travail régulier.

En second lieu, nous étudions un système de files d'attente M/M/c avec vacances actives différenciées, interruptions de vacances actives selon une loi de Bernoulli, waiting servers, et clients impatients (balking et reneging). Ce modèle introduit le concept de vacances actives différenciées, comprenant des vacances longues, prises après avoir servi au moins un client (vacances actives de type 1), et des vacances courtes, prises immédiatement après le retour des vacances précédentes avec un système vide (vacances actives de type 2). Nous obtenons la solution à l’état stable, formulons une fonction de coût et effectuant une analyse coût-revenu-profit.

Mots clés: Modèles de files d'attente Markoviens, vacances actives, mesures de performance, clients impatients, modèle de coûts.

Abstract :

The main contribution of this thesis is the introduction of novel Markovian queueing models that capture the complex dynamics of call center operations. These models use the working vacation policy, which allows servers to perform secondary tasks at reduced service rates when the system is idle, a common and effective strategy in call centers, where servers can handle different types of calls or other activities during idle periods.

The main objectives of this research encompass deriving steady-state probabilities and performance measures for these queueing systems, utilizing the matrix geometric method. Additionally, we develop cost models and conduct sensitivity analyses to comprehensively assess the impact of various system parameters on performance metrics.

At first, we examine an M/M/c queue with waiting servers, balking, reneging, and K-variant working vacations, subject to Bernoulli schedule vacation interruption. The steady-state probabilities of the system are obtained, and the cost function is optimized by a direct search method to find the optimal service rates during during working vacation and regular working periods.

Then, we investigates an M/M/c queueing system with differentiated working vacations (DWV), Bernoulli schedule working vacation interruption, waiting servers, and customer impatience (balking and reneging). This model introduces the concept of differentiated vacations, consisting of long vacations, taken after serving at least one customer (type-1 vacation), and short vacations, taken immediately after returning from the previous vacation with an empty system (type-2 vacation). We obtain the steady-state solution, formulate a cost function and perform cost-revenue-profit analysis.

Keywords: Markovian queueing models, working vacation, performance measures, customer impatience, cost model.