N° d'ordre :

Université de Saida– Dr. Moulay Tahar
Faculté Mathématiques, Informatique et Télécommunications

Thèse

Présentée pour obtenir le diplôme de

Doctorat 3ème Cycle

Spécialité : Mathématiques Appliquées.

Filière : Mathématiques Appliquées

Par :

Yatim Nada Riheb

Thème :

Analysis of unreliable machining systems

Thèse soutenue le 06 Juillet 2026 devant le jury composé de :

N°	Nom et prénom	Grade	Etablissement	Qualité
01	Madani Fethi	Prof.	Université de Saida – Dr. Moulay Tahar	Président
02	Bouchentouf Amina Angelika	Prof.	Université de Sidi Bel Abbes- Djillali Lyabes	Rapporteur
03	Vijaya Laxmi Pikkala	Prof.	Université de Andhra, Visakhapatnam, Inde	Co-rapporteur
04	Mechab Boubeker	Prof.	Université de Sidi Bel Abbes- Djillali Lyabes	Examinateur
05	Rahmani Saadia	Prof.	Université de Saida – Dr. Moulay Tahar	Examinateur
06	Yahiaoui Lahcene	MCA	Université de Saida – Dr. Moulay Tahar	Examinateur

Dedication

I dedicate this work to everyone who supported me along the way.

To my father and mother, who believed in me and saw in me the ability to succeed, and to my dear siblings who provided me with emotional support at all times.

And I do not forget my classmates, who shared this journey with me and enriched this scientific and moral path. From them, I learned the basics of work and elevated myself by dealing with them.

And to everyone who contributed, from near or far, in supporting this work until its realization.

Acknowledgements

I thank and praise ALLAH Almighty for guiding me, his grace in granting me the path of knowledge and making it accessible. Praise be to ALLAH, for without His guidance, this humble work would not have been completed.

I would like to extend my sincere thanks and gratitude to all those who contributed to the completion and realization of this work and put it on sound scientific foundations. Foremost among them is my respected supervisor, Professor Bouchentouf Amina Angelika, who spared no effort or knowledge to accomplish it on the right track, by providing continuous support and valuable advice during all stages of research and study.

I do not forget the great contribution made by my co-supervisor, Professor Pikkala Vijaya Laxmi in achieving the desired results during the research process.

I would also like to express my sincere appreciation to the members of my committee for their interest in reviewing and evaluating this thesis.

Finally, I thank the the esteemed professors of Dr. Moulay Tahar university of Saida, for providing the supportive environment and resources needed to accomplish this work.

الملخص

تطور هذه الأطروحة نماذج عشوائية لتحليل أداء وتحسين الأنظمة المعرضة للأعطال وعدم صبر العملاء (الرفض والتراجع)، من خلال نمذجة البيئات غير الموثوقة التي قد تتعرض فيها الخدمة للانقطاع أو التدهور، تضع هذه الدراسة أساسًا لتقييم التصنيع الصناعي والاتصالات اللاسلكية.

أولاً، ندرس نظام التصنيع الآلي (AMS) المعرض للكوارث وفترات تدهور الخدمة (أعطال العمل) وإجازات الخادم (مفردة ومتعددة)، يتم دمج عدم صبر العملاء (الرفض والتراجع)، لتقدير خسائر النظام. يتم نمذجة النظام كطابور M /M /c/N. باستخدام طريقة التحليل المصفوفي (MAM)، نستنتج مقاييس الأداء في حالة الاستقرار، ونجري تحليل التكلفة والعائد، ونطبق مقاربة التحسين الذي يجمع بين البحث المباشر وطرق شبه نيوتن لتحديد العدد الأمثل للخوادم وسعة النظام ومعدلات الخدمة خلال فترات التشغيل العادية والإصلاح.

بناءً على هذا الأساس التحليلي، يتناول الجزء الثاني شبكات IEEE 802.16e اللاسلكية من خلال نمذجة آلية توفير الطاقة (PSM) كطابور M /M /1. يدمج هذا النموذج سياسات الإجازة المتباينة، وخادم الانتظار، ونفاد صبر العملاء اعتماداً على الحالة والأحداث الكارثية، إضافة إلى فترات تدهور الخدمة. يتم استخدام طريقة المصفوفة الهندسية (MGM) لاشتقاق شروط الاستقرار وتوزيعات الاحتمالات لحالة الاستقرار، ونتبع بتقديم نموذج التكلفة لتقييم الأثر الاقتصادي لتنفيذ (PSM). يسلط التحليل العددي الضوء على كيفية تأثير معلمات النظام على الأداء.

توفر نتائج هذه الأطروحة أدوات عملية لتصميم أنظمة مرنة تحافظ على الكفاءة والجدوى الاقتصادية على الرغم من الاضطرابات التشغيلية.

كلمات مفتاحية: نظرية الطوابير، طرق تحليل المصفوفات (MAM)، طريقة هندسة المصفوفات (MGM)، أنظمة التصنيع غير الموثوقة، نفاد صبر العملاء، إجازات الخادم، التحسين.

Abstract

This thesis develops stochastic models to analyze the performance and optimization of systems subject to failures and customer impatience (balking and reneging). By modeling unreliable environments where service may be disrupted or degraded, the research establishes a foundation for evaluating industrial manufacturing and wireless telecommunications.

First, we study an Automatic Manufacturing System (AMS) exposed to disasters, degraded service periods (working breakdowns), and server vacations (single and multiple). Customer impatience, balking and reneging, is incorporated to quantify system losses. The system is modeled as an $M/M/c/N$ queue. Using the matrix–analytic method (MAM), we derive steady–state performance measures, conduct a cost–benefit analysis, and apply an optimization approach combining direct search and quasi–Newton methods to determine the optimal number of servers, system capacity, and service rates during both normal and repair periods.

Building on this analytical foundation, the second part addresses IEEE 802.16e wireless networks by modeling the Power Saving Mechanism (PSM) as an $M/M/1$ queue. This model integrates differentiated vacation policies, a waiting server, state–dependent customer impatience, catastrophic events, and degraded service periods. The matrix–geometric method (MGM) is employed to derive stability conditions and steady–state probability distributions, complemented by a cost model to evaluate the economic impact of PSM implementation. Numerical analysis highlights how system parameters influence performance.

The results of this thesis provide practical tools for designing resilient systems that sustain efficiency and economic viability despite operational disruptions.

keywords: Queueing theory, matrix–analytic methods (MAM), matrix–geometric method (MGM), unreliable manufacturing systems, customer impatience, server vacations, optimization.

Résumé

Cette thèse est consacrée au développement de différents modèles stochastiques pour analyser les performances et l'optimisation de systèmes soumis à des défaillances et à l'impatience des clients (dérobade (balking) et abandon (reneging)). En modélisant des environnements non fiables où le service peut être interrompu ou dégradé, cette étude fournit une base fondamentale pour l'évaluation des systèmes de production industrielle et des télécommunications sans fil.

Dans un premier temps, nous étudions un système de fabrication automatisé (AMS) soumis à des désastres, à des pannes opérationnelles (où le service continue mais à un taux réduit) et à des périodes de vacances (simples et multiples). Le modèle intègre également un serveur en attente (waiting server) ainsi que l'impatience des clients (balking et reneging), afin de quantifier la perte de capacité du système. Le système est modélisé par une file d'attente $M/M/c/N$. En utilisant la méthode d'analyse matricielle (Matrix–Analytic Method, MAM), nous dérivons les mesures de performance en régime permanent, menons une analyse coût–bénéfice et appliquons une approche d'optimisation combinant la recherche directe et la méthode quasi-Newton, afin de déterminer le nombre optimal de serveurs, la capacité optimale du système, ainsi que les taux de service optimaux pendant les périodes normales et de panne.

En nous appuyant sur cette base analytique, l'étude s'étend aux réseaux sans fil IEEE 802.16e en modélisant le mécanisme d'économie d'énergie (PSM) par une file d'attente $M/M/1$. Ce modèle intègre des politiques de vacances différenciées, waiting server, l'impatience des clients dépendante de l'état, des événements catastrophiques et des pannes opérationnelles. La méthode géométrique matricielle (Matrix–Geometric Method, MGM) est utilisée pour établir les conditions de stabilité et les distributions de probabilité à l'équilibre, puis un modèle de coût est développé pour évaluer les implications économiques de la mise en œuvre du PSM. Une analyse numérique approfondie révèle l'impact des différents paramètres sur les performances du système.

Les résultats de cette thèse offrent des outils pratiques pour la conception de systèmes résilients, capables de maintenir leur efficacité et leur viabilité économique en dépit des perturbations opérationnelles.

Mots clés : Théorie des files d'attente, méthode d'analyse matricielle (MAM), méthode géométrique matricielle (MGM), systèmes de production non fiables, impatience des

clients, vacances du serveur, optimisation.

List of works

List of research papers

1. Yatim, Nada Riheb & Bouchentouf, Amina Angelika & Vijaya Laxmi, Pikkala. A queueing model for an automatic manufacturing system with disasters, breakdowns and vacations. Optimal design and analysis. Operations Research and Decisions.Vol. 34, No. 4 (2024) | DOI: 10.37190/ord240413

2. Bouchentouf, Amina Angelika & Vijaya Laxmi, Pikkala & Yatim, Nada Riheb. Performance analysis of power saving mechanisms in mobile wireless communication: A queueing model with vacations, impatience, and disruptions, submitted.

List of presentations

1. Yatim, Nada Riheb, Statistical analysis of a single-server Markovian feedback queueing model with balking. The International Workshop on Statistics and its Applications (IWSA'23). March 01 to 02, 2023. University Dr. Moulay Taher of Saida, Algeria.

2. Yatim, Nada Riheb & Bouchentouf, Amina Angelika."Customer Behavior in Queueing Models : Analyzing Single-Server System with Working Vacations, Interruptions, and Impatience". M2DS'24 National Conference: Mathematical Modeling for Dynamic Systems, June 26 to 27, 2024. Freres Mentouri Constantine 1 university , Algeria.

3. Yatim, Nada Riheb. "Analysis of Queueing models with Impatience". Algerian Student Week Celebration May 19. May 18 to 22, 2025. Dr.Moulay Tahar Saida University. Algeria.

4. Yatim, Nada Riheb & Bouchentouf, Amina Angelika & Vijaya Laxmi, Pikkala. "A queueing model for an automatic manufacturing system with disasters, breakdowns and vacations. Optimal design and analysis. Operations Research and

Decisions". The 2nd International Conference on Recent Advances Mathematics and Informatics IEEE-ICRAMI 2025, November 20 to 23, 2025, Sousse, Tunisia.

Contents

List of Figures

List of Tables

Introduction

Queueing systems describe the interaction between two main entities: the customer (which can be a person, a machine, or a data packet) and the server (the service provider). When demand increases, queues form, especially if many arrivals occur at the same time, leading to delays. Initially, interest in queueing theory emerged from telecommunications research after 1917, notably with Erlang's work on telephone networks. Since then, the application of queueing models has expanded from healthcare and traffic systems to more complex structures in computer science, operations research, and machining systems.In a manufacturing context, the arrival of parts for assembly creates a queueing system where machine speed and product quality are critical to avoid production stops. The main goal of studying these systems is to reduce congestion, sometimes called queue explosion, by ensuring efficient service. Any failure or disaster in a manufacturing facility slows productivity and causes economic loss, which shows the need for strong mathematical models.Servers may also become unavailable during vacations, such as maintenance or idle periods. These breaks, while necessary to keep equipment working longer, increase queue length. They may also cause customers to leave the queue (reneging) or decide not to join it (balking).Theoretical analysis allows for the evaluation of these systems under various conditions. This thesis utilizes Matrix-Analytic Methods (MAM) and the Matrix-Geometric Method (MGM) to provide closed-form expressions for steady-state probabilities. These tools are particularly effective for Quasi-Birth-and-Death (QBD) processes, where transition rates are modulated by an underlying Markov chain.

Problem statement

Unreliable queueing systems are common in modern manufacturing. However, most studies examine factors such as server vacations, breakdowns, or customer impatience separately. This fails to capture the real complexity of industrial and communication systems. While impatience has often been studied in human systems (banks, airports), it is equally important in manufacturing, where delays directly increase costs. This thesis addresses this gap by building integrated models that include disasters, vacations, and impatience together.

Research objectives

The main objective is to provide mathematical solutions for realistic models of unreliable manufacturing systems. Specifically, we aim to:

- Develop complex queueing models that integrate sudden failures and maintenance phases.
- Employ MAM and MGM to extract steady-state performance measures.
- Conduct cost-benefit optimizations to assist in decision-making regarding system capacity and service rates.

Contributions of the thesis

This work proposes new mathematical models that combine server unreliability with vacation policies and customer behavior. By providing analytical results for integrated systems, it offers practical guidance for managing workflow, such as improving maintenance planning and reducing customer loss. In this way, the research connects theoretical modeling with real industrial practice.

Thesis outline

- Chapter 1 presents the fundamentals of queueing models, including Poisson, Markov, and PH renewal processes, as well as the theoretical foundations of the Matrix-Analytic Method (MAM) and the Matrix-Geometric Method (MGM).
- Chapter 2 provides a comprehensive literature review of queueing models incorporating server vacations and customer impatience. It places a special focus on unreliable systems, thereby identifying the specific scientific gaps that this thesis aims to bridge.
- Chapter 3 investigates an Automated Manufacturing System (AMS) incorporating disasters, working breakdowns, waiting servers, customer impatience, and vacations. This system is analyzed as an $M/M/c/N$ queue to evaluate performance and optimize total cost. This chapter is based on the paper: "A queueing model for an automatic manufacturing system with disasters, breakdowns, and vacations: optimal design and analysis", published in Operations Research and Decisions, 34(4), 231-250, (2024) [113].
- Chapter 4 focuses on IEEE 802.16e wireless networks by modeling the Power Saving Mechanism (PSM) as an $M/M/1$ queueing system. This chapter derives stability conditions and evaluates the economic impact of the mechanism under various network dynamics.

1 Theoretical background and Matrix-Analytic foundations

1.1 Fundamentals of queueing theory

1.1.1 Basic queueing system characteristics

The classification of diverse queueing models is based on their fundamental components. The identification of these components is the key to optimal modeling and analyzing queueing problems. Queueing systems are typically represented as a process in which customers arrive to receive service and may wait if the servers are occupied. The fundamental components of these systems are classified as follows:

- The customer arrival process: Customer arrivals are described as random variables that express the number of arrivals (arrival rate) or the time between two consecutive arrivals (inter-arrivals time). In standard models, these inter-arrival times are assumed to be independent and identically distributed.

- Service process: This refers to the operations provided by the server to the customers. It is typically characterized by the service time required for each customer, which is considered a sequence of independent and identically distributed (i.i.d.) random variables. These follow a specific probability distribution such as:

  - Exponential: The simplest Markovian case with a constant service rate.
  - Erlang: A multi-stage process with sequential phases.
  - General (G): A non-specific distribution typically used for robustness analysis.
  - Phase-Type (PH): A highly versatile class of distributions that represent the time until absorption in a Markov chain. Due to its ability to approximate any non-negative distribution and its structured matrix form, the PH distribution serves as the cornerstone of the Matrix-Analytic approach employed in this thesis.

- The source of the population: The set of potential customers from which clients emerge (i.e., clients, machines, jobs, etc.). It is classified as either finite or infinite.

- Service discipline: It specifies how the server selects the next customers to serve. However, there are several possibilities for the order in which customers will be served. The main service disciplines are as follows:

  - FCFS (First-Come First-Served): This system is the most common. Customers are served in order of arrival. (ex: Customer service office in a bank branch.)
  - FIFO (First In First Out): The term is employed chiefly in the context of the storage or processing sequence of items within a structure. Thus, the item that is placed first into the buffer is retrieved first. (ex: The processing of patients' blood samples in a hospital emergency unit)
  - LIFO (last in, first out): The last customer in line is the first to be served.
  - SIRO (Service in Random Order): The next customer to be served is randomly chosen from the queue(ex: A high-speed network router is a device that handles a buffer of data packets requiring transmission through the same link.)
  - Priority: Customers waiting for service are sorted and selected according to their urgency levels rather than strictly on their arrival time. They are usually classified as Preemptive vs. Non-Preemptive, where a high-priority arrival can interrupt a service already in progress.
  - PS (Processor Sharing): Customers are served equally,the capacity of the system is shared among the customers (ex: Data Networks: Bandwidth Sharing among multiple user while downloading data through a Wi-Fi router).

1.1.1.1 Kendall's notations

To classify the various queueing systems, the notation described initially by Kendall is utilized, which has a specific form consisting of six symbols arranged in the order $A/B/c/m/n/Z$, to describe the main characteristics of queues models where:

- A: describe the distribution of inter-arrival times between two successive customers.

- B: describe the distribution of service times.
  To specify the distributions A and B, the following symbols are introduced:

  - M: inter-arrivals of customers are independent and identically distributed according to an exponential distribution. It corresponds to a point Poisson process (memoryless property).

- E_k: This symbol denotes a process in which the time intervals between two successive arrivals are independent and identically distributed random variables following an Erlang distribution of order k.
- H_k: hyper-exponential distribution of degree k.
- D: The inter-arrival times of customers or the service times are constant and always the same.
- G: Customer inter-arrival times have a general distribution and may be dependent.
- GI: The inter-arrival are independent and identically distributed and times follow a general distribution.

- c: is the number of parallel servers (single, multiple or infinite).
- m: the system capacity is the total number of clients allowed in the entire system, including those on the waiting list and those on the servers.
- n: The size of the source population of customers.
- Z: service discipline which specifies the order by which jobs in the queue are being served (FCFS, FIFO, LIFO...etc.)

1.1.2 Performance measures

The primary objective of analyzing complex manufacturing systems is the extraction of key performance metrics under various operational scenarios. These metrics provide the quantitative foundation required to evaluate system efficiency and facilitate informed economic decision-making. Furthermore, a rigorous assessment of performance measures allows for the optimization of system design, ensuring that the modeled infrastructure can effectively meet the challenges of real-world industrial environments.

1.1.2.1 System state metrics

These metrics represent the fundamental steady-state outputs of the model. They characterize the customer flow, the average occupancy of the system, and the efficiency of the service process.

- Mean system size (L_s): The average number of customers present in the system (both waiting and in service).
- Effective arrival rate (λ_{eff}): The actual rate at which customers join the system, accounting for those lost due to balking.
- Mean waiting time (W_s): The average time a customer spends in the system. This is typically related to L_s via Little's Law: $L_s = \lambda_{eff} W_s$.

- Mean throughput (E_s): The average number of customers successfully served per unit of time.

1.1.2.2 Server availability and operational modes

These measures quantify the proportion of time the service facility spends in various states. They are instrumental in balancing service capacity against operational costs across different stochastic regimes:

- Probability of regular busy period (P_B): The system is operating at full service capacity μ.
- Probability of working vacation (P_{WV}): The server provides service at a reduced rate μ_v while on vacation.
- Probability of complete vacation (P_V): The server is completely unavailable for service.
- Probability of repair (P_R): The system is inactive due to a breakdown and is undergoing a recovery process.

1.1.2.3 Customer impatience and loss metrics

These metrics quantify the economic and operational leaks caused by poor service quality or system interruptions. They serve as critical inputs for cost-optimization functions:

- Average reneging rate (R_{ren}): The rate at which customers, after initially joining the queue, leave the system before receiving service due to excessive waiting time. This behavior is typically modeled by a reneging probability or a hazard rate that depends on the current state of the system and the customer's patience threshold, which may vary depending on whether the server is active, on vacation, or under repair.
- Average balking rate (R_{bal}): The rate at which potential customers decide not to join the queue upon arrival. This behavior is typically modeled by a joining probability b_n, which depends on the number of customers n already in the system or the current state of the server (e.g., whether it is on vacation or broken down).
- Probability of loss (P_{loss}): The overall probability that a potential customer is not served, combining the effects of balking, reneging, and disasters.

1.1.2.4 System reliability and resilience

Reliability metrics provide a predictive measure of service continuity. In manufacturing contexts, these indicators are essential for developing maintenance schedules and mitigating the impact of catastrophic failures:

- System reliability (R): Defined as the probability that the system performs its required function without failure under stated conditions for a specified period.

- Availability (A): The probability that the system is in a functioning state (busy or working vacation) at any given time t.

To provide a clear overview of the metrics discussed above, Table 1.1 summarizes the fundamental performance indicators and their associated notations. These parameters constitute the primary outputs for the steady-state analysis of the queueing models developed in the subsequent chapters of this thesis.

Table 1.1: Summary of fundamental performance metrics and notations.

Symbol	Description	Unit/Type
L_s	Mean number of customers in the system	Customers
W_s	Mean waiting time in the system	Time
λ_{eff}	Effective arrival rate	Cust./Unit time
P_{bal}	Average balking rate	Cust./Unit time
R_{ren}	Average reneging rate	Cust./Unit time
P_{loss}	Overall probability of customer loss	Probability
P_B, P_{WV}, P_V, R	Probabilities of operational states	Probability
A	System availability	Probability

Remark 1.1.1. *The notations presented in this chapter serve as a general framework for the fundamental metrics of queueing systems. In the subsequent chapters, these symbols will be further refined and indexed (e.g., R_{ren} or P_{bal}) to account for specific system configurations involving diverse vacation policies, breakdown scenarios, and multi-phase service regimes.*

1.2 Markov processes and stochastic counting processes

1.2.1 Markov process

Markovian queues are among the earliest models studied in queueing theory, grounded in the theory of Markov processes initiated by the Russian mathematician Andrey Markov in 1907. Unlike classical probability where for a sequence of events, the future outcome is assumed to be independent of the present, this new concept expresses the random event's correspondence at a specific moment in time with its previous state. We consider a Markov process as a stochastic process defined as follows:

Definition 1.2.1. *(Markov Process). Let $\{X(t) : t \geq 0\}$ be a stochastic process defined on the time index set $\mathcal{T} = [0, \infty)$, where $X(t)$ represents the state of the process at time t. For any sequence of times $0 \leq t_1 < t_2 < \cdots < t_n$, and for any states in the state space $\mathcal{S}$, a stochastic process is a Markov process if it satisfies:*

$$P[X(t_n) \leq x_n \mid X(t_{n-1}) = x_{n-1}, \ldots, X(t_1) = x_1] = P[X(t_n) \leq x_n \mid X(t_{n-1}) = x_{n-1}]. \quad (1.1)$$

Equation (1.1) signifies the memoryless property of a Markov process, indicating that the future states of a Markov process depend only on the current states, known as the Markov property.

1.2.2 Continuous time Markov chains (CTMCs)

We assume that our stochastic process satisfies the Markov property for a set of discrete states over continuous time.

Definition 1.2.2. *(CTMC). Let* $\{X(t) : t \geq 0\}$ *be a stochastic process in continuous-time* $\mathscr{T}$ *on a discrete state space* $\mathscr{S}$*. It is a continuous-time Markov chain if for all* $s, t \geq 0$ *and all* $i, j, i_r \in \mathscr{S}$*:*

$$P(X(t+s) = j \mid X(s) = i, X(r) = i_r, 0 \leq r < s) = P(X(t+s) = j \mid X(s) = i).$$

A CTMC is considered time-homogeneous if the transition probability between two states depends only on the duration of the interval t, rather than the specific starting point s:

$$P(X(s+t) = j \mid X(s) = i) = P(X(t) = j \mid X(0) = i) = p_{ij}(t),$$

$\forall s, t \geq 0$ and $i, j \in \mathscr{S}$. The set of all such probabilities forms the transition matrix $\mathbf{P}(t) = [p_{ij}(t)]$, which satisfies the Chapman-Kolmogorov equations [23]:

$$p_{ij}(s+t) = \sum_{k \in S} p_{ik}(s) p_{kj}(t), \quad \forall s, t \geq 0.$$

1.2.2.1 The infinitesimal generator matrix Q

The local dynamics of a CTMC are fully characterized by its transition rates, which form the infinitesimal generator matrix Q. For $i \neq j$, the rate q_{ij} is defined as the derivative of the transition probability at the origin [88]:

$$q_{ij} = p'_{ij}(0) = \lim_{\Delta t \to 0} \frac{P(X(\Delta t) = j \mid X(0) = i)}{\Delta t}.$$

Physically, q_{ij} represents the frequency of jumps from state i to j. The diagonal elements q_{ii} (often denoted as $-q_i$) are defined such that the row sums of Q are zero:

$$q_{ii} = -q_i = -\sum_{j \neq i} q_{ij}.$$

This structure implies that the *sojourn time* (waiting time) in state i is an exponentially distributed random variable with mean $1/q_i$. The infinitesimal generator Q takes the following block form:

$$Q = \begin{bmatrix} -q_0 & q_{01} & q_{02} & \cdots \\ q_{10} & -q_1 & q_{12} & \cdots \\ q_{20} & q_{21} & -q_2 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

This matrix governs the evolution of the transition probability matrix $P(t)$ through the *Kolmogorov forward equations*:

$$\frac{d}{dt}P(t) = P(t)Q.$$

1.2.3 Steady-state analysis

For the complex queueing models discussed in this thesis (involving vacations and failures), we are primarily interested in the long-run behavior of the system. If the CTMC is irreducible and positive recurrent, the transition probabilities $p_{ij}(t)$ converge to a stationary distribution $\boldsymbol{\pi} = (\pi_0, \pi_1, \dots)$ as $t \to \infty$.
Then, let $Q = \{q_{ij}\}$ be the infinitesimal generator of the CTMC $\{X(t), t \geq 0\}$ on the state space $\mathscr{S}$. For a stationary probability distribution $\{\pi_i\}_{i\in\mathscr{S}}$, where $\pi_i \geq 0$ for all $i \in \mathscr{S}$, we have the following result:

Proposition 1.2.1. *[75] The stationary probability vector $\boldsymbol{\pi} = (\pi_0, \pi_1, \dots)$ is the unique solution to the system:*

$$\boldsymbol{\pi} Q = \mathbf{0}, \quad \textit{subject to} \quad \boldsymbol{\pi}\mathbf{e} = 1, \tag{1.2}$$

where $\mathbf{e}$ *is a column vector of ones.*

The relation (1.2) shows that CTMC reaches equilibrium, where we obtain the steady-state probabilities indicating that the rate of entry into a state is equal to the rate of exit, with a firm probability transition.

1.3 Phase-type (PH) distributions

1.3.1 Continuous phase type (CPH) distribution

This section is based on Breuer and Baum [22]. Consider an irreducible CTMC, $\{Y(t) : t \geq 0\}$, with K transient states and one absorbing state, $K+1$, with generator Q defined as:

$$Q = \begin{bmatrix} T & T^0 \\ \mathbf{0} & 0 \end{bmatrix}, \tag{1.3}$$

where T is an $K \times K$ matrix with $T_{ii} < 0$ and $T_{ij} \geq 0$ for $i \neq j$, governing the transitions among the transient states. T^0 is a column vector of order K (the exit vector) that governs the transitions from transient states to the absorbing state, such that:

$$T\mathbf{e} + T^0 = \mathbf{0}, \tag{1.4}$$

which follows from the property that the row sums of an infinitesimal generator matrix must equal zero.

The initial distribution of the CTMC is the row vector $\tilde{\boldsymbol{\alpha}} = (\boldsymbol{\alpha}, \alpha_{K+1})$, where $\boldsymbol{\alpha}$ is a row vector of dimension K. Usually, in queueing models, we assume $\alpha_{K+1} = 0$, meaning the process starts in a transient state.

Let $X := \inf\{t \geq 0 : Y(t) = K+1\}$ be the random variable representing the time until absorption.
The distribution of X is called a phase-type distribution (PH distribution) with parameters $(\boldsymbol{\alpha}, T)$, denoted by $X \sim PH(\boldsymbol{\alpha}, T)$. The dimension K is the *order* of the distribution, and the transient states $\{1, \ldots, K\}$ are referred to as *phases*.

The basic characteristics of a PH distribution are:

- The cumulative distribution function: $F(t) = 1 - \boldsymbol{\alpha} e^{Tt}\mathbf{e}, \quad t \geq 0,$
- The density function: $f(t) = \boldsymbol{\alpha} e^{Tt} T^0, \quad t \geq 0,$
- The n^{th} moment: $K_n = (-1)^n n! \boldsymbol{\alpha} T^{-n}\mathbf{e}$.

Proofs of these formulas can be found in the lecture notes of He [44].

1.3.1.1 Some specific PH-distributions

In this part, a few special phase-type distributions used in this thesis are defined with their fundamental expressions [23].

- Exponential: The classical exponential distribution with parameter λ is a scalar case of CPH distribution where:

$$\boldsymbol{\alpha} = (1), \quad T = (-\lambda).$$

  The fundamental expressions are:

$$\begin{aligned} f(t) &= \lambda e^{-\lambda t}, \quad t \geq 0, \\ F(t) &= 1 - e^{-\lambda t}, \quad t \geq 0, \\ m_n &= \frac{n!}{\lambda^n}, \quad n \geq 1. \end{aligned}$$

- Erlang: The Erlang distribution of order K consists of K sequential phases, each with rate λ:

$$\boldsymbol{\alpha} = (1, 0, \ldots, 0), \; T = \begin{bmatrix} -\lambda & \lambda & & \\ & -\lambda & \lambda & \\ & & \ddots & \lambda \\ & & & -\lambda \end{bmatrix}_{K \times K}.$$

  Fundamental expressions:

$$\begin{aligned} f(t) &= \frac{\lambda^K t^{K-1} e^{-\lambda t}}{(K-1)!}, \quad t \geq 0, \\ m_1 &= \frac{K}{\lambda}, \quad \sigma^2 = \frac{K}{\lambda^2}. \end{aligned}$$

- Generalized Erlang: Also known as the hypoexponential distribution, it consists of K phases with different rates $\lambda_1, \ldots, \lambda_K$:

$$\boldsymbol{\alpha} = (1, 0, \ldots, 0), \quad T = \begin{bmatrix} -\lambda_1 & \lambda_1 & & \\ & -\lambda_2 & \lambda_2 & \\ & & \ddots & \lambda_{K-1} \\ & & & -\lambda_K \end{bmatrix}.$$

- Hyperexponential: A finite mixture of K exponentials with mixing probabilities α_k:

$$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K), \; T = \mathrm{diag}(-\lambda_1, -\lambda_2, \ldots, -\lambda_K).$$

  The density and distribution functions are:

$$f(t) = \sum_{k=1}^{K} \alpha_k \lambda_k e^{-\lambda_k t}, \quad t \geq 0,$$
$$F(t) = 1 - \sum_{k=1}^{K} \alpha_k e^{-\lambda_k t}, \quad t \geq 0.$$

1.3.1.2 Counting process

Definition 1.3.1. *(Counting process). A stochastic process $\{N(t), t \geq 0\}$ is called a counting process if $N(t)$ represents the total number of events that have occurred up to time t. It must satisfy:*

1. *$N(t) \geq 0$,*
2. *$N(t)$ is an integer for all $t \geq 0$,*
3. *If $s \leq t$, then $N(s) \leq N(t)$ (non-decreasing).*

For $s < t$, the difference $N(t) - N(s)$ represents the number of events that occurred during the interval $(s, t]$. This process is fundamental in queueing theory to model, for instance, the cumulative number of arrivals to a system over time.

1.3.1.3 Poisson process

The Poisson process is the most widely used counting process in queueing theory, typically modeling the arrival process of customers.

Definition 1.3.2. *(Poisson process). A counting process $\{N(t), t \geq 0\}$ is a Poisson process with rate $\lambda > 0$ if it satisfies the following conditions:*

- *Stationary increments: The probability of having k events in an interval depends only on the length of the interval t, and not on its starting point s:*

$$P(N(t+s) - N(s) = k) = P(N(t) = k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!}, \quad k = 0, 1, \ldots.$$

- *Independent increments: The number of events occurring in disjoint time intervals are independent random variables. For $s < t$:*

$$P(N(t+s) - N(s) = k, N(s) = j) = P(N(t+s) - N(s) = k)P(N(s) = j).$$

- *Infinitesimal behavior: As $\Delta t \to 0$, the probability of transitions is given by:*

$$P(N(t+\Delta t) - N(t) = k) = \begin{cases} 1 - \lambda\Delta t + o(\Delta t) & \text{if } k = 0 \\ \lambda\Delta t + o(\Delta t) & \text{if } k = 1 \\ o(\Delta t) & \text{if } k \geq 2. \end{cases}$$

The parameter λ represents the arrival rate or intensity, signifying the average number of events per unit of time.

1.3.1.4 Renewal process

The renewal process is a generalization of the Poisson process where the inter-arrival times $X_1, X_2, \ldots$ are independent and identically distributed (i.i.d.) random variables following any distribution F. In this context, an event is called a *renewal*.

Let $\{X_n, n \geq 1\}$ be the sequence of i.i.d. inter-arrival times with $P[X_n > 0] = 1$. The arrival times (or renewal times) are defined by the partial sums:

$$S_0 = 0, \quad S_n = \sum_{k=1}^{n} X_k, \quad n \geq 1.$$

Definition 1.3.3. *The process $\{N(t) : t \geq 0\}$ defined by:*

$$N(t) = \sum_{n=1}^{\infty} \mathbb{1}_{\{S_n \leq t\}} = \max\{n : S_n \leq t\}$$

is called a renewal process.

Example 1.3.1. *To illustrate the concept, consider a system where "inter-renewal times" X_n represent the time required to repair a machine. When a machine breaks down, it is immediately sent for repair. S_n represents the completion time of the n-th repair, and $N(t)$ counts the total number of machines repaired up to time t.*

1.3.1.5 PH-Renewal process

A Poisson process is a special case of a renewal process where the inter-arrival times are exponentially distributed. In the context of phase-type distributions, this corresponds to a PH-distribution of order $K = 1$.

Definition 1.3.4. *(PH-Renewal process). A continuous-time renewal process $\{X_n, n \geq 1\}$ is a CPH-renewal process if the inter-renewal times follow a CPH distribution with an irreducible representation $(\boldsymbol{\alpha}, T)$ of order K.*

This generalization is fundamental for modeling real-world systems that exhibit non-exponential inter-arrival times (such as bursty traffic or multi-stage service processes). By using PH-distributions, the renewal process maintains a Markovian structure, which is justified by the following property:

Proposition 1.3.1. *[22] (Conditional memoryless property). For a phase-type distribution $PH(\boldsymbol{\alpha}, T)$, the remaining time Z until absorption, given that the current phase is i, is expressed by:*

$$P(Z \leq t \mid \text{current phase} = i) = 1 - \mathbf{e}_i e^{Tt}\mathbf{e},$$

which corresponds to a $PH(\mathbf{e}_i, T)$ distribution, where $\mathbf{e}_i$ denotes the i-th canonical row vector.

This property implies that the remaining time until an arrival (absorption) depends only on the current phase i. Consequently, the CPH-renewal process can be modeled as a two-dimensional CTMC, $\{N(t), J(t)\}$, on the state space $\mathscr{S} = \{(i, j) : i \geq 0, 1 \leq j \leq K\}$, where $N(t)$ denotes the number of arrivals up to time t, and $J(t)$ represents the current phase. The generator matrix $\tilde{Q}$ takes the following block-upper-triangular form:

$$\tilde{Q} = \begin{bmatrix} T & T^0\boldsymbol{\alpha} & \mathbf{0} & \cdots \\ \mathbf{0} & T & T^0\boldsymbol{\alpha} & \cdots \\ \mathbf{0} & \mathbf{0} & T & \ddots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

The condition $T\mathbf{e} + T^0\boldsymbol{\alpha}\mathbf{e} = \mathbf{0}$ ensures that the row sums of $\tilde{Q}$ are zero. The dynamics of the process are described as follows:

- Transitions without renewal: These represent phase changes within the same level (no arrival). These internal transitions are governed by the sub-generator matrix T.
- Transitions with renewal: These occur when the process reaches the absorbing state and immediately restarts in a new phase according to the initial probability vector $\boldsymbol{\alpha}$. This transition from level i to $i+1$ is governed by the product matrix $T^0\boldsymbol{\alpha}$.

This Markovian description is essential for applying Matrix-Analytic Methods (MAM) to solve complex queueing systems such as $PH/PH/1$, $M/PH/1$, and $M/PH/c$.

1.4 Matrix-Analytic Methods and QBD processes

We attribute the reputation of Matrix Analytic Methods (MAM) as one of the most widely used methods for queueing analysis to its ability to analyze a large category of complex Markovian models applied in engineering, communications, manufacturing, and technology.

Marcel Neuts [76] provided the first assessment of the MAM concept to the scientific community in the mid-seventies. This was followed by remarkable progress, notably by Latouche and Ramaswami [65], which introduced the analytical methods required to study a variety of complex real-world waiting situations. These developments have been further extended and applied in recent works such as those by Chakravarthy et al. [24], Assar [8], Ziad et al. [118], and Chettouf et al. [27]. These contributions have revolutionized the field of queueing theory by establishing algorithmic probability as a crucial instrument for solving theoretical and practical queueing problems, as evidenced by the seminal works of Neuts [76] and Latouche and Ramaswami [65].

MAM is a powerful tool for studying the stationary distribution of an ergodic Quasi-Birth-and-Death (QBD) process. This is particularly effective when the stationary distribution satisfies the matrix-Geometric property, allowing it to be expressed in the following form:

$$\boldsymbol{\pi}_n = \boldsymbol{\pi}_0 R^n, \quad n \geq 0,$$

where $\boldsymbol{\pi}$ is the unique solution to the system $\boldsymbol{\pi} Q = \mathbf{0}$ subject to the normalization condition $\boldsymbol{\pi}\mathbf{e} = 1$. The stationary distribution vector $\boldsymbol{\pi} = [\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots]$ is partitioned into sub-vectors (or phase-vectors) $\boldsymbol{\pi}_i = [\pi_{i,1}, \pi_{i,2}, \ldots, \pi_{i,K}]$, where K denotes the number of phases in each level.

1.4.1 Quasi Birth-and-Death processes (QBD)

The process of Quasi-Birth-and-Death (QBD) is a type of birth-and-death process whose transition rates are modulated by an underlying continuous Markov chain. A QBD process is a mathematical model used to describe systems with a multi-dimensional state space where states are grouped into "levels" and "phases", which constitutes the key structure for the matrix-Analytic approach. We begin by providing a simple, intuitive foundation: the Birth-Death process, and then assert the definition of the QBD process.

Definition 1.4.1. *(Birth-Death Process) A continuous-time Birth-Death process is a Markov chain $\{X_t, t > 0\}$ with values on the state space $\Omega = \{0, 1, 2, ...\}$, with the following generator matrix Q:*

$$Q = \begin{pmatrix} -b_0 & b_0 & 0 & 0 & \ldots \\ b_2 & -(b_0 + b_2) & b_0 & 0 & \ldots \\ 0 & b_2 & -(b_0 + b_2) & b_0 & \ldots \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix},$$

where b_0 is the birth rate and b_2 is the death rate, with $b_0, b_2 > 0$.

The Birth-Death process is the scalar case of a QBD process with single-element blocks, in which the system changes by either increasing its state by one (a "birth") or decreasing its state by one (a "death"). This is a common model in queueing theory; it perfectly represents the $M/M/1$ queue, where arrivals are referred to as births and departures as deaths.

Assuming the process is non-explosive, the limiting probability is defined as $\pi_i = \lim_{t\to\infty} P[X_t = i]$.

Theorem 1.1. *[65] The stationary distribution of the homogeneous Birth-Death process satisfies:*

$$\pi_i = \pi_{i-1}\rho, \quad i \geq 1,$$

where $\rho = b_0/b_2$. The process is ergodic if and only if $b_0 < b_2$.

Theorem 1.2. *[23] Under the condition that $b_0 < b_2$, with both of them being positive, the steady-state probability vector $\boldsymbol{\pi} = \{\pi_0, \pi_1, \pi_2, \ldots\}$ of Q such that $\boldsymbol{\pi} Q = \mathbf{0}$ and $\boldsymbol{\pi}\mathbf{e} = 1$ is given by*

$$\pi_n = (1-\rho)\rho^n, \quad n \geq 0.$$

In the same manner, for a QBD process, we follow a similar approach to obtain the matrix-geometric form; however, the scalar ρ is replaced by the rate matrix R.

Definition 1.4.2. *(QBD Process). Let $\{(X(t), Y(t)) \mid t \geq 0\}$ be a continuous-time Markov process on the state space:*

$$\Omega = \{(i, j) : i \geq 0, 1 \leq j \leq K\}, \tag{1.5}$$

which we partition into levels $l(i) = \{(i,1),(i,2),\ldots,(i,K)\}$. The first coordinate i is called the level, and the second coordinate j is the phase. A Markov chain is called a QBD if transitions are restricted to the same level or to the two adjacent levels: a transition from (i, j) to (i', j') is possible only if $i' \in \{i-1, i, i+1\}$.

Definition 1.4.3. *(Homogeneous QBD Process). A Markov chain is a homogeneous QBD if the transition rates between levels n and $n \pm 1$ do not depend on the value of n for $n \geq 1$. The infinitesimal generator Q takes the block tri-diagonal form:*

$$Q = \begin{pmatrix} B_0 & A_0 & \mathbf{0} & \mathbf{0} & \cdots \\ B_1 & A_1 & A_0 & \mathbf{0} & \cdots \\ \mathbf{0} & A_2 & A_1 & A_0 & \cdots \\ \mathbf{0} & \mathbf{0} & A_2 & A_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}, \tag{1.6}$$

where A_0, A_1, A_2 are square matrices of order K. The row sums of Q are zero, implying:

- $(B_0 + A_0)\mathbf{e} = \mathbf{0}$
- $(B_1 + A_1 + A_0)\mathbf{e} = \mathbf{0}$
- $(A_2 + A_1 + A_0)\mathbf{e} = \mathbf{0}$

1.4.2 Matrix-Geometric Method for analyzing QBD process

Quasi-birth-death (QBD) processes may be analyzed through the Matrix Geometric Method which is a powerful tool to study complex multi-dimensional systems. This part is devoted to demonstrating the application of the MGM for the QBD process, as outlined in the relevant chapter (10.6) of Stewart [94].

Let Q be the infinitesimal generator of a QBD process, and the stationary distribution is obtained by solving $\boldsymbol{\pi} Q = \mathbf{0}$. Let $\boldsymbol{\pi}$ be partitioned conformally with Q into sub-vectors:

$$\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots), \tag{1.7}$$

where $\boldsymbol{\pi}_i = (\pi(i,1), \pi(i,2), \ldots, \pi(i,K))$. Neuts [76] built the Matrix Geometric solutions for QBD under the assumption that the process is positive recurrent, hence the ergodicity of the QBD process is affirmed in the following result:

Theorem 1.3. *[65](Drift condition) If the QBD is irreducible, if K is finite, and if the stochastic matrix $A = A_0 + A_1 + A_2$ is irreducible, then the QBD is positive recurrent if and only if:*

$$\mu = \boldsymbol{\alpha} A_0 \mathbf{e} < \boldsymbol{\alpha} A_2 \mathbf{e},$$

where $\boldsymbol{\alpha}$ is the stationary probability vector of A.

The steady state equations are as follow:

$$\boldsymbol{\pi}_0 B_0 + \boldsymbol{\pi}_1 B_1 = \mathbf{0}, \tag{1.8}$$

$$\boldsymbol{\pi}_0 A_0 + \boldsymbol{\pi}_1 A_1 + \boldsymbol{\pi}_2 A_2 = \mathbf{0}, \tag{1.9}$$

$$\boldsymbol{\pi}_{i-1} A_0 + \boldsymbol{\pi}_i A_1 + \boldsymbol{\pi}_{i+1} A_2 = \mathbf{0}, \quad i \geq 2. \tag{1.10}$$

Equations (1.8)-(1.10) possess a repetitive structure. Hence there exists a constant matrix R such that a recursive solution for the stationary probability vector $\boldsymbol{\pi}_i$ could be determined.

Theorem 1.4. *[65](Matrix-Geometric property) Assume that the continuous time QBD process is positive recurrent. Then, the stationary probability distribution is such that:*

$$\boldsymbol{\pi}_{i+1} = \boldsymbol{\pi}_i R, \quad i \geq 1, \tag{1.11}$$

where the matrix R records the rate of sojourn in the states of $l(i+1)$ per unit of the local time of $l(i)$. Furthermore, we have that $R = A_0 N$, where the matrix N records the expected sojourn time in the states of $l(i)$, starting from $l(i)$, before the first visit to $l(i-1)$.

Given $\boldsymbol{\pi}_0$, $\boldsymbol{\pi}_1$ and R, the rest of the stationary distribution sub-vectors $\boldsymbol{\pi}_i$ can be extracted using the following theorem.

Theorem 1.5. *[76] The process Q is positive recurrent if and only if the minimal nonnegative solution R to the matrix-quadratic equation:*

$$A_0 + R A_1 + R^2 A_2 = \mathbf{0}, \tag{1.12}$$

has all its eigenvalues inside the unit disk and the finite system of equations:

$$\boldsymbol{\pi}_0(B_0 + RB_1) = \mathbf{0},$$
$$\boldsymbol{\pi}_0(I - R)^{-1}\mathbf{e} = 1, \tag{1.13}$$

has a (unique) positive solution $\boldsymbol{\pi}_0$.
If the matrix A is irreducible, then $sp(R) < 1$ *if and only if* $\boldsymbol{\alpha} A_2\mathbf{e} > \boldsymbol{\alpha} A_0\mathbf{e}$, *where* $\boldsymbol{\alpha}$ *is the stationary probability vector of A. The stationary probability vector* $\boldsymbol{\pi} = [\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \ldots]$ *of Q is given by:*

$$\boldsymbol{\pi}_i = \boldsymbol{\pi}_0 R^i, \quad \textit{for } i \geq 0. \tag{1.14}$$

By replacing (1.14) into equation (1.10), we obtain:

$$\boldsymbol{\pi}_1 R^{i-2} A_2 + \boldsymbol{\pi}_1 R^{i-1} A_1 + \boldsymbol{\pi}_1 R^i A_0 = \mathbf{0}, \quad i \geq 2,$$

which implies $\boldsymbol{\pi}_1 R^{i-2}(A_2 + RA_1 + R^2 A_0) = \mathbf{0}$. Since A_1 is invertible, we can compute R numerically using the iterative algorithm developed by Neuts [76]. By following successive substitutions, it results from (1.12):

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

such that $R_0 = \mathbf{0}$, noting that R_k converges to R. The algorithm is stopped when $\|R_{k+1} - R_k\| < \epsilon$, where ϵ is a specified tolerance criterion. This algorithm is considered weak due to the long computation times and the large number of iterations required. Consequently, Latouche and Ramaswami [65] introduced a more efficient algorithm by identifying two matrices G and U:

- The matrix G is the solution to: $G = (-A_1)^{-1}A_2 + (-A_1)^{-1}A_0G^2$. The term $(-A_1)^{-1}A_2$ gives the probability that the QBD process moves from level $l(n)$ to $l(n-1)$; the second term $(-A_1)^{-1}A_0G^2$ represents moving from $l(n)$ to $l(n+1)$ and eventually returning to $l(n-1)$ via level $l(n)$.

- We define $U = A_1 + A_0G$. U is the generator of a transient process starting in level $l(n)$ and restricted to that level until the first visit to level $l(n-1)$.

Lemma 1.1. *[65] The matrices R, G and U are related as follows:*

$$\begin{aligned}
U &= A_1 + A_0G \\
G &= (-A_1)^{-1}A_2 + (-A_1)^{-1}A_0G^2 \\
R &= A_0(-U)^{-1} \\
U &= A_1 + RA_2 \\
R &= -A_0(A_1 + A_0G)^{-1} \\
G &= -(A_1 + RA_2)^{-1}A_2.
\end{aligned}$$

1.4.2.1 Logarithmic reduction algorithm

The following steps describe the computation of the matrices R and G using the logarithmic reduction algorithm for the continuous-time case. Given the matrices A_0, A_1, and A_2 for the ergodic generator, and a small positive tolerance ε, the algorithm proceeds as follows (see [65] for full details):

Step 0: $J \leftarrow (-A_1)^{-1}A_0$, $L \leftarrow (-A_1)^{-1}A_2$, $G = L$, and $T = J$.

Step 1: Carry out the following computations in the specified order:

1. $V = JL + LJ$
2. $M = J^2$
3. $J \leftarrow (I - V)^{-1}M$
4. $M \leftarrow L^2$
5. $L \leftarrow (I - V)^{-1}M$
6. $G \leftarrow G + TL$
7. $T \leftarrow TJ$

Step 2: If $\|\mathbf{e} - G\mathbf{e}\|_\infty > \varepsilon$, return to **Step 1**. Otherwise, proceed to **Step 3**.

Step 3: Having computed G, the rate matrix is obtained by:

$$R = -A_0(A_1 + A_0G)^{-1}.$$

To complete the stationary distribution, we solve the boundary equations. Based on the structure of the infinitesimal generator Q, the first two equations of the system $\boldsymbol{\pi}Q = \mathbf{0}$ are:

$$\boldsymbol{\pi}_0 B_0 + \boldsymbol{\pi}_1 B_1 = \mathbf{0}, \tag{1.15}$$

$$\boldsymbol{\pi}_0 A_0 + \boldsymbol{\pi}_1 A_1 + \boldsymbol{\pi}_2 A_2 = \mathbf{0}. \tag{1.16}$$

By substituting the Matrix-Geometric relation $\boldsymbol{\pi}_2 = \boldsymbol{\pi}_1 R$ into Equation (1.16), we obtain the reduced system:

$$\boldsymbol{\pi}_0 A_0 + \boldsymbol{\pi}_1 (A_1 + RA_2) = \mathbf{0}.$$

Combined with (1.15) and the normalization condition $\pi_0 + \boldsymbol{\pi}_1(I - R)^{-1}\mathbf{e} = 1$, these linear equations allow for the unique determination of the initial vectors π_0 and $\boldsymbol{\pi}_1$.

Example: The $M/PH/1$ queueing model

The $M/PH/1$ queue model is a fundamental application of the QBD process. In this system, arrivals follow a Poisson process with rate λ, and service times are governed by a PH-distribution $(\boldsymbol{\alpha}, T)$ of order K.

The process $\{(X(t), Y(t)), t \geq 0\}$ is a CTMC on the state space $\Omega = \{0\} \cup \{(i, j) : i \geq 1, 1 \leq j \leq K\}$, where $X(t)$ is the number of customers and $Y(t)$ is the current service phase. The transitions are structured as follows:

- Arrivals (Birth): An arrival increases the level from i to $i+1$. If the system was empty ($i=0$), the service starts in a phase chosen according to $\boldsymbol{\alpha}$.
- Service completions (Death): A service completion (absorption in the PH-process) decreases the level from i to $i-1$. If $i-1>0$, a new service begins immediately with initial vector $\boldsymbol{\alpha}$.
- Phase changes: Transitions within the same level i occur as the customer progresses through service phases without completing the service.

The infinitesimal generator Q matches the QBD structure previously defined:

$$Q = \begin{bmatrix} -\lambda & \lambda\boldsymbol{\alpha} & \mathbf{0} & \mathbf{0} & \dots \\ \mathbf{T}^0 & T-\lambda I & \lambda I & \mathbf{0} & \dots \\ \mathbf{0} & \mathbf{T}^0\boldsymbol{\alpha} & T-\lambda I & \lambda I & \dots \\ \mathbf{0} & \mathbf{0} & \mathbf{T}^0\boldsymbol{\alpha} & T-\lambda I & \dots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{bmatrix}. \tag{1.17}$$

By identification with the homogeneous QBD parameters, we have:

- Boundary blocks: $B_0 = -\lambda$ and $B_1 = \mathbf{T}^0$.
- Transition blocks ($n \geq 1$): $A_0 = \lambda I$, $A_1 = T - \lambda I$, and $A_2 = \mathbf{T}^0\boldsymbol{\alpha}$.

For this specific model, the Matrix-Geometric rate matrix R has an explicit closed-form solution, which bypasses the need for iterative algorithms.

Theorem 1.6. *[76] We have*

$$\begin{aligned} \pi_0 &= 1-\nu, \\ \boldsymbol{\pi}_i &= (1-\nu)\boldsymbol{\alpha}R^i, \quad i \geq 1. \end{aligned}$$

The matrix R is given by $R = \lambda(\lambda I - \lambda\mathbf{e}\boldsymbol{\alpha} - T)^{-1}$, where $\nu = \lambda(-\boldsymbol{\alpha}T^{-1}\mathbf{e}) < 1$, such that $(-\boldsymbol{\alpha}T^{-1}\mathbf{e})$ is the mean service time.

This mathematical framework provides the necessary tools to analyze more complex systems. In chapters 3 and 4, we will extend these results to queueing models incorporating server vacations and breakdowns.

2 Literature review

We dedicate this part to examine the theoretical study of unreliable servers through previous research in contrast to the modern challenges faced by current systems, so that we can identify the questions that the models in this thesis aim to solve.

2.1 Review of queueing model features

This section reviews the main features of queueing models that are relevant to this thesis. The subsections highlight specific aspects such as breakdowns, vacations, and customer impatience, which are later integrated into the proposed models.

2.1.1 Disruptions in queueing systems

In real-world operational environments, servers are rarely perfectly reliable. They are subject to various disruptions that critically impair service delivery, transitioning the system from a classical steady-state model to a more complex unreliable server framework. These disruptions reduce performance, inflate operational costs, and degrade system reputation.

2.1.1.1 Categories of server disruptions

Disruptions can be classified into three fundamental categories based on their stochastic impact on the service rate μ:

1. Server breakdowns (complete failure): The service rate drops to zero ($\mu = 0$) during the interruption, and the server remains unavailable until a repair process is completed.

    - Temporary Breakdowns: Short-term malfunctions where the server returns to operation after a reset or minor repair.
      Example: In computing systems, a CPU overheating that triggers an automatic reboot or a temporary software freeze requiring a system restart.

- Permanent breakdowns: Irrecoverable failures where the server unit must be replaced or undergoes extensive overhaul.
  Example: A hardware failure such as a burnt-out motherboard in a data center or a terminal mechanical fracture in a manufacturing lathe.

2. Working breakdowns (partial failure): The server remains operational but at a degraded capacity ($0 < \mu_b < \mu$). Service is still provided, but at a significantly lower speed or efficiency.
   Example: In communication systems, software bugs or memory leaks might not crash the server entirely but cause a decline in packet processing speed. In industrial production, a machine operating in "safe mode" due to a faulty sensor continues to produce units but at a reduced output rate.

3. Catastrophes (system clearing): High-impact events that affect the entire system state. Unlike a simple breakdown, a catastrophe often clears the current queue or forces a total system reset.
   Example: A power interruption in a plastic injection molding line. The immediate cooling causes the resin to harden inside the molds, requiring a complete system flushing (removal of damaged parts and cleaning) before the manufacturing process can be reset.

2.1.1.2 Timing of disruptions: Operational states

The impact of a failure is highly dependent on the operational phase of the server at the moment the disruption occurs:

- Breakdown during busy period: The failure occurs while a customer is being served. This introduces specific service disciplines, such as service-preemptive-resume (continuing from the point of interruption after repair) or *Service-Preemptive-Repeat* (restarting the service from the beginning).

- Breakdown during Vacation (or working vacation): Even during working vacations (where the server operates at a lower rate μ_v), the system remains susceptible to failures. A breakdown in this state is particularly critical as it transitions the system from a low-speed state to a total halt, often acting as a catalyst for increased customer impatience.

2.1.1.3 Dynamics of the repair process

The mathematical complexity of the solution is dictated by the statistical nature of the recovery phase:

- Exponential repair distribution: If the repair process is memoryless (rate η), the system can be modeled using pure Markovian processes and solved via the matrix-analytic method (MAM), preserving algorithmic stability.

- General repair distribution (G): In cases where repairs have a fixed minimum duration or variable complexity, the process is non-Markovian. This necessitates the introduction of a supplementary variable (representing the elapsed repair time), justifying the use of the supplementary variable technique (SVT) to restore the Markovian property.
- Recovery policies: Repairs may be immediate or delayed based on a threshold policy (e.g., N-policy), where repair only begins once the queue reaches a specific length, adding a strategic optimization dimension to the reliability analysis.

2.1.1.4 Socio-economic and operational impacts

Studying these disruptions allows for the design of robust strategies to mitigate negative consequences:

- Performance degradation: Addressing task accumulation and long wait times resulting from frequent downtime.
- Output diminishment: Evaluating the reduction in total system throughput and its impact on product quality.
- Economic implications: Balancing the financial aspects of repair costs and production losses against declining customer satisfaction and reputation damage.

2.1.2 Literature review on system unreliability

Recently, a diverse range of stochastic models has been implemented to assess the impact of server failure on system performance. This body of work primarily addresses scenarios where server malfunctions lead to a complete operational shutdown. For instance, Gupta and Kumar [41] presented a retrial $M/M/1$ system prone to failures and repair periods, incorporating an infinite capacity orbit and multiple working vacations to optimize performance metrics under low service-rate conditions. Similarly, Peter and Mary [80] evaluated a single-server system with heterogeneous arrivals, accounting for breakdowns occurring during both the busy period and working vacation phases.

The methodological approach to these problems often involves matrix-analytic techniques. In Assar [8], the matrix geometric method (MGM) was employed to conduct a comparative stability study of automated teller machine (ATM) systems, providing key performance indicators to quantify the cost of server failure. Furthermore, Bouchentouf et al. [13] utilized a finite Markovian model to represent multi-station machining systems, integrating Bernoulli feedback and customer impatience to establish a robust sensitivity analysis. The foundational understanding of these failure-prone systems has been significantly enriched by various researchers, including Afroun et al. [3], Atencia and Moreno [9], Bansal and Moses [11], Kumar and Jain [64], Shin [93], Towsley and Tripathi [104], among others.

2.1.2.1 Working breakdowns and fault-tolerant systems

A more nuanced perspective considers cases where the server continues to provide service during malfunctions, albeit at a reduced rate, a phenomenon known as a working breakdown. This approach is vital for modeling resilient infrastructures. For example, Kumar and Jain [64] analyzed the reliability of Fault-Tolerant Systems (FTS) utilizing standby machines, while Deena Merit and Haridass [30] investigated an $M^X/G/1$ queueing system facing disasters with the support of a substitute server

The integration of specific management policies during these periods has also been a focal point of recent studies. Recently, Bouchentouf et al. [21] modeled and optimized an automated manufacturing system (AMS) incorporating differentiated vacation (DV) policies, waiting servers, and impatient customers under failure conditions. Additionally, Yang et al. [112] addressed cost optimization in finite-capacity Markovian systems, emphasizing the importance of setup times following a working breakdown. Further, Varshney et al. [106] analyzed machining systems by considering the simultaneous impact of working breakdowns and repairman availability.

2.1.2.2 Catastrophic failures and system recovery

Extreme disruptions, or *catastrophes*, where the entire system state is reset, pose the most significant challenge to operational stability. Early research by Kumar and Arivudainambi [63] provided transient solutions for the system size in Markovian queues subject to catastrophes. Expanding on this, Demircioglu et al. [32] studied the impact of disaster occurrences and the subsequent system reset phase in discrete-time $GI/GI/1$ systems. Recent advancements by Ramdani et al. [85] proposed finite-capacity multi-server queues with feedback and catastrophes, offering a comprehensive analysis of both transient and steady-state behaviors within a fluid queue framework.

2.1.2.3 Synthesis of reliability and recovery strategies

These collective studies reflect the critical challenges faced by modern machining and telecommunication systems, where frequent and unexpected failures lead to production delays, reduced quality, and escalated costs. The existing literature by Bansal and Moses [11], Bouchentouf et al. [13, 21], Chakravarthy et al. [24], Kumar and Jain [64], Varshney et al. [106], and Ramdani et al. [85] underscores the necessity of integrating multi-stage repair processes to ensure operational continuity.

Beyond standard repair mechanisms, advanced interventions such as standby support are essential. As explored by Baby Saroja and Suvitha [10], Chakravarthy et al. [24], and Deena Merit and Haridass [30], auxiliary servers act as a strategic response to prevent customer loss during breakdowns. In the specific context of telecommunications, as highlighted by Xu et al. [110], Zaghouani et al. [116], and Chettouf et al. [27], the primary challenge lies in minimizing the recovery window, the sum of the system restart time and the refill time required to return to a state of stability. Optimizing these parameters is key to mitigating the damages caused by catastrophic disruptions like cyber-attacks or power outages.

2.1.3 Server vacations in queueing systems

In modern manufacturing and service environments, operational efficiency is increasingly tied to the economic management of resources. The monitoring of server performance relative to operational costs has led to the development of *server vacation* models. These models represent periods where the server is intentionally made unavailable for primary tasks due to several factors: the absence of customers, scheduled maintenance, energy-saving protocols, or the execution of secondary auxiliary tasks.

The conceptual basis of the vacation policy lies in the trade-off between system idleness and resource availability. While an idle server contributes to increased operational overhead without generating output, a complete shutdown risks losing incoming traffic or increasing wait times beyond acceptable thresholds. Vacation policies were thus introduced to balance these concerns, providing a realistic framework for modeling industrial systems [100, 101]. For a comprehensive overview of early developments in vacation queues, we refer to the classic surveys by Tian et al. [103], Ke et al. [59], and Fiems [38].

Various vacation disciplines have been established in the literature to suit different operational requirements:

- Single vacation policy: The server takes exactly one vacation at the end of each busy period, returning to wait for arrivals regardless of the system state at the end of that vacation, [53, 70, 73].

- Multiple vacation policy: The server continues to take successive vacations until it finds at least one customer waiting upon its return, [70, 73].

- Exhaustive service discipline: The server serves all customers until the system is completely empty before initiating a vacation, [1, 96].

- Gated vacation queues: A gate is placed upon the server's return; only customers present at that specific instant are served before the next vacation [86, 97].

- K-variant vacation queues: The server is permitted to take up to K sequential vacations if the system remains empty, after which it must return to the busy period, [16, 20, 118] .

- Differentiated vacation queues:In this case, the server has the option to take two types of vacations. When the server served at least one customer, it has the option to enter the first type of vacation; upon returning and finding the system empty, the server immediately enters the second type of vacation,[18, 47, 98, 105].

2.1.3.1 The working vacation (WV) policy

A significant evolution in this field is the *Working Vacation* (WV) policy. Unlike traditional models where service completely ceases, the WV assumption allows the server to continue processing tasks at a reduced rate ($\mu_v < \mu$) during the vacation period. This principle is highly practical for manufacturing systems as it alleviates server stress during low-demand periods without completely halting the workflow.

The foundational model for the working vacation principle was introduced by Servi and Finn [90]. Since then, research has expanded significantly into both single and multiple working vacation policies, as seen in the works of Wu and Takagi [109] and Majid and Manoharan [70, 71]. Recent literature continues to refine these models through optimization and stationary solutions, including contributions from Panta et al. [78], Ziad et al. [118], Divya and Indhira [35], and Ramdani et al. [84].

2.1.3.2 Integration of vacation periods in unreliable systems

A critical analysis of the current literature reveals that server vacations and system failures are frequently treated as independent phenomena. This simplification often fails to capture the complexity of industrial reality, where disruptions can occur regardless of the server's operational state.

Several researchers have begun addressing this integration:

- Karthick and Suvitha [58] and Jeyakumar [53] investigated systems vulnerable to catastrophic failures and disasters within vacation frameworks, specifically in manufacturing contexts.

- Baby Saroja and Suvitha [10] examined the stability of an $M/Hyper-Erlang/1$ queue combining single working vacations with breakdowns and standby support.

- Peter and Mary [80] analyzed a single-server system with heterogeneous arrivals, highlighting the impact of breakdowns occurring during both the busy and working vacation periods.

Despite these advancements, there is a notable lack of models that account for the simultaneous impact of multiple vacation cycles, complex failure modes, and customer impatience within a single unified framework. The omission of these coupled dynamics renders current models less effective for representing the high-risk, multi-state environments found in modern production systems.

2.1.4 Impatience in queueing models

The analysis of customer behavior is a fundamental pillar of modern queueing theory, as it transitions mathematical models from purely mechanical arrival-service processes to human-centric or resource-sensitive systems. In contemporary industrial and service environments, the efficiency of a system is not solely measured by its throughput, but also by its ability to retain and satisfy its incoming workload.

In systems prone to disruptions, such as manufacturing facilities facing frequent server failures, the interaction between service interruptions and customer behavior becomes critical. When a system halts, either partially or completely, the resulting accumulation of tasks or customers often leads to what is defined as impatience phenomena. From an industrial perspective, this accumulation can trigger a storage crisis, where physical or digital buffer capacities are exceeded, potentially leading to product damage or data loss. From a human or service perspective, prolonged delays result in a profound loss of trust in the service entity's reliability.

Mathematically, these factors necessitate the integration of state-dependent departure rates to account for the stochastic nature of customer reneging. If the effects of impatience are not mitigated through optimal service policies, they lead to substantial revenue losses and compromise the system's long-term stability. Therefore, characterizing the specific mechanisms of impatient behavior, such as balking and reneging, is essential for developing robust operational strategies. These behaviors are typically categorized as follows:

- Balking: When the system is not considered idle and the customer arrive at service unit, we consider him as balked costumer when he decide not to join the line at all [16, 42, 89, 114, 115].

- Reneging : This phenomena is defined as joining a line but leaving without being served,we explain as follow: after joining the line the costumer decide to leave without being served the time spend is superior than his available waiting time [43, 62, 114, 115].

- Retrial: While often studied as a distinct stochastic process, the retrial phenomenon is fundamentally driven by customer impatience. When a customer encounters an occupied or unreliable server, their immediate impatience may lead to balking or reneging. However, instead of exiting the system permanently, the customer may join a virtual pool known as an *orbit*. From this orbit, the customer re-applies for service after a random duration. In this context, retrials represent a bridge between impatient behavior and system persistence [4, 41, 91].

- Jockeying: This behavior represents a form of dynamic impatience specific to multi-server systems. It is defined as the movement of a customer from one queue to another, perceived to be moving faster or having a shorter waiting line, in an attempt to minimize total waiting time. Mathematically, jockeying acts as a self-organizing mechanism that balances the workload across servers, though it significantly increases the dimensionality of the state-space analysis [37, 82].

2.1.4.1 Evolution of research on impatient behavior

The foundational interest in customer attitudes within queueing systems dates back to the classical literature, where Barrer [12] and Haight [43] first established the mathematical basis for analyzing reneging and balking. Building upon these seminal works, a significant body of advanced research has emerged to address more complex scenarios. Contemporary contributions, such as those by Ammar [5], Wang et al. [108], Bouchentouf et al. [14], Bouchentouf et al. [15], Suranga Sampath and Jicheng [98], and Sharma et al. [92], have successfully integrated these behavioral traits into various stochastic environments, reflecting the increasing need for realistic system modeling.

Despite this progress, the behavioral aspects of customers, specifically balking and reneging, are often addressed in a simplified manner in current literature, frequently treated as phenomena isolated from the dynamics of system failures. However, recent studies have begun to bridge this gap by examining the interplay between impatience and server unreliability.In the context of severe disruptions, Zhang and Gao [117] and Ammar et al. [6] explored $M/M/1$ queues subject to disasters and working breakdowns, focusing on how catastrophic events that remove all present customers affect system performance. Kuki et al. [61] investigated finite-capacity retrial systems ($M/M/1/N$), incorporating the joint effects of server failures, impatience, and customer collisions. Additionally, Dehamnia et al. [31] derived ergodicity conditions for retrial queues by considering varying levels of customer patience.

While these works provide valuable insights, there remains a persistent need for a steady-state multidimensional analysis that fully integrates unpredictable failures with the specific, state-dependent impatience patterns of customers in infinite-capacity production systems. This thesis aims to address these limitations by providing a more holistic modeling approach.

2.2 Mathematical methodologies in existing literature

The analysis of unreliable queueing systems requires determining the state probabilities $\pi_{n,j}$ to extract performance measures, where n represents the number of customers and j denotes the server's state (e.g., normal, vacation, breakdown). The complexity of the chosen mathematical approach often acts as a bottleneck in providing realistic and stable solutions for high-dimensional models.

2.2.1 Probability Generating Functions (PGF) and Transform Methods

In classical queueing theory, the PGF method is used to transform a system of discrete difference equations into a continuous functional equation. For a sequence of probabilities $\{P_n\}_{n=0}^{\infty}$, the PGF is defined as:

$$P(z) = \sum_{n=0}^{\infty} P_n z^n, \quad |z| \leq 1$$

For transient (time-dependent) analysis, the Laplace Transform is typically applied:

$$\mathscr{L}\{P_n(t)\} = \int_0^\infty e^{-st} P_n(t) dt$$

Recent literature has utilized these tools extensively. Jeyakumar [53] used the supplementary variable technique to obtain PGF functions for various parameters in production line models. Similarly, Muthukumar and Bagyam [72], Suvitha and Ohapriyadharsini [99] utilized PGFs to evaluate systems with working vacations, disasters, and balking, while Ramdani et al. [85] employed Laplace transforms to solve for transient behaviors in Markovian queues.

Limitations in multi-state and complex systems: While effective for basic models, these approaches face significant structural challenges in higher dimensions:

- Analytical intractability: In systems with multiple server states (e.g., breakdowns, various vacations, and recovery phases), the scalar $P(z)$ becomes a *vector* of functions. Solving the resulting coupled equations requires identifying the roots of high-degree characteristic polynomials (often via Rouché's Theorem), a process that is mathematically exhaustive and prone to error.
- Multidimensional explosion: The multi-state nature of modern systems, where a server might simultaneously be in a vacation period while prone to failure, renders the PGF approach nearly impossible to apply. The complexity is driven by the number of auxiliary variables needed to describe the system's phase space.
- Numerical instability: Extracting individual probabilities P_n from $P(z)$ through successive differentiation or contour integration becomes numerically unstable for large n.

2.2.2 Recursive method

The recursive method is an alternative numerical approach used to determine steady-state probabilities by expressing P_n as a function of P_{n-1} or P_0. For instance, Ahuja and Jain [4] analyzed a queueing system with an unreliable service provider, retrial policy, and delayed threshold recovery using a recursive scheme.

Limitations in multi-state systems:

- Computational complexity: In multi-state systems, the recurrence is no longer a simple scalar relation; it becomes a system of linear equations for each level n. As the number of phases increases, the computational cost grows quadratically.
- Numerical error propagation: Because each P_n depends on the precision of P_{n-1}, rounding errors can accumulate rapidly, particularly in systems with infinite capacity or low service rates.

2.2.3 Supplementary variable technique (SVT)

When queueing systems involve non-exponential distributions (such as general service times G), the process $X(t)$ is non-Markovian. To restore the Markov property, a continuous supplementary variable $X_s(t)$, representing the elapsed or remaining service time, is introduced. This leads to a system of partial differential equations (PDEs), as seen in the works of Jeyakumar [53] and Dehamnia et al. [31].

Limitations: The application of SVT to multidimensional systems presents a dimensionality curse. If a model includes multiple non-exponential processes (e.g., general service and general repair times), one must add multiple supplementary variables, leading to high-dimensional PDEs that are often analytically intractable.

2.2.4 Transform methods

Transform methods convert differential equations into algebraic ones. Ramdani et al. [85] and Varshney et al. [106] used these for transient solutions and unreliable systems. However, the primary drawback is the inversion problem. Finding the analytical inverse $\mathscr{L}^{-1}\{P_n^*(s)\}$ is rarely possible for multi-state models, and numerical inversion algorithms (e.g., Stehfest) are highly sensitive to the precision of the poles, often producing non-physical results near stability boundaries.

2.2.5 The Matrix-Analytic Method (MAM)

To overcome the limitations of the PGF, SVT, and Transform methods, specifically the difficulty of solving high-dimensional PDEs and the numerical risks of inversion, this thesis utilizes the Matrix-Analytic Method (MAM).

MAM provides stable and computationally efficient solutions for multidimensional problems by treating the queueing process as a Quasi-Birth-Death (QBD) process. This approach preserves the physical transition structure of the system in block-matrix form. It offers an algorithmic path to the steady-state solution $\pi_n = \pi_1 R^{n-1}$, where R is the rate matrix. This is particularly vital for the infinite-dimensional production systems studied herein, which integrate unpredictable failures, complex vacation cycles, and customer impatience patterns.

3 A queueing model for an automatic manufacturing system with disasters, working breakdowns, and vacations: optimal design and analysis

3.1 Introduction

This chapter focuses on the study of an unreliable machining system modeled as a continuous-time $M/M/c/N$ queueing model with single and multiple vacation policies, waiting servers, disasters, repair, working breakdowns, reneging, and balking. Continuous-time queues have proven to be powerful tools for analyzing the performance of various systems. Their broad range of applications and versatility has made them a key area of research in fields such as computer science, communication systems, and manufacturing [28, 29, 39, 87]. Disasters, also called queue flushing [104] or mass exodus [26] break down the server and clear the system of all work. However, they have no effect when the system is empty. These models can be applied to computer networks or industrial systems that are vulnerable to virus infections or reset failures. Extensive literature exists on this subject (e.g.,[9, 32, 49, 55, 67, 111]). Another aspect of queueing models with disasters is the possibility of working breakdowns. These are situations where the service does not stop completely but slows down due to a server malfunction. For example, a computer virus may reduce the system's performance without shutting it down. Many service sectors, such as transportation, telecommunications, and healthcare, encounter this problem when a backup server takes over the main server until it is repaired. Kalidass and Kastouri [57] pioneered this concept in queueing systems, and [60] applied it to a $M/G/1$ queue with disasters and working breakdown services. Later [7], they studied an $M/G/1$ preemptive priority retrial queue with the same features. Recently, Kim ans Lee [30] explored state-dependent arrival and optional re-service in an $M^X/G/1$ queue with disasters and working breakdown service. Vacation queues are an important class in which the server may take a break once the system gets empty.

These queues are widely studied in queueing theory, as they can be used to improve the efficiency of many real-world systems, such as call centers, healthcare facilities, industry, and transportation. Some of the early notable works on this topic were presented in [36, 100–102]. A significant amount of papers have been devoted to the study of customers' impatience in different contexts of vacation queueing models. Customers may balk, meaning they do not join the queue if it is too long or for other reasons. They may also renege, meaning they leave the queue before being served if they wait too long. These phenomena have been well discussed in the case of vacation, where the server does not serve any customers during the vacation period (cf.[3, 5, 16, 77, 98, 115]) and in the case of working vacation; the server acts at a slow rate of service during the vacation of the server (cf. [13–15, 17, 46, 62, 118]). This chapter proposes a new $M/M/c/N$ queueing model that incorporates various features and adaptations to capture the dynamics of real-world systems, such as multiple and single vacation policies, disasters, working breakdowns, and impatience (balking and reneging). The queueing system has potential applications in different industrial settings, such as manufacturing plants and assembly lines. It combines several characteristics that have not been studied together in the literature. This makes the model complex and challenging to analyze. We employ the Q-matrix method, a powerful numerical technique for analyzing the steady-state behavior of Markov chains, to conduct the steady-state analysis of the model. We calculate the steady-state probabilities and derive closed-form expressions for various performance measures of the queueing system. It is worth pointing out that there are different techniques for solving mathematical problems, such as analytical approaches that use probability-generating functions or maximum entropy approaches, and matrix-analytic approaches that compute the stationary distribution of Markov chains. In our chapter, we use the Q-matrix method because it is suitable for complicated queueing problems that lack straightforward analytical solutions. This method simplifies the construction of system transition probability matrices and expresses the steady-state probabilities as functions of specific state probabilities. We also construct a cost model and formulate an optimization problem to determine the optimal operating values of different system parameters such as the system capacity, the number of servers, and the service rates during repair and regular busy periods, that minimize the total expected cost of the system. We use the direct search method and the quasi-Newton method to solve the optimization problem.

The remainder of the chapter is organized as follows. Section 3.2 describes the queueing model. Section 3.3 represents the practical example of the proposed queueing model. Section 3.4 derives the steady-state distribution of the queueing system. Section 3.5 determines different performance measures. Section 3.6, on the other hand, develops a cost model and offers numerical results. In Section 3.7, we conclude the chapter.

3.2 The mathematical description of the model

An $M/M/c/N$ unreliable multi-server queueing system has been considered. Different assumptions required for the formulation of the model are as:

- The system has c servers that provide service to incoming customers who arrive independently according to Poisson processes with rate λ. The service times follow i.i.d. exponential distributions with parameter $1\mu_B$ and are independent of the arrivals. When a customer arrives and sees that one of the c servers is free, he is immediately served according to the first-come-first-served (FCFS) discipline. Otherwise, the customer has to wait for his turn to receive service.

- After serving all customers, the servers enter a waiting period before going on vacation. The waiting period follows an exponential distribution with parameter ϖ. Once this period ends, the servers go on vacation for a random duration following an exponential distribution with rate ϕ. The system operates in two vacation modes: Multiple vacation (MV) and Single vacation (SV). In MV mode, if there are no waiting customers at the end of the vacation period, another vacation period begins. Otherwise, the system enters a service phase. On the other hand, in SV mode, when the vacation period ends, the servers switch to the busy period and remain idle until a new arrival occurs. The Kronecker δ is given as:

$$\delta = \begin{cases} 1, & \text{for the single vacation model,} \\ 0, & \text{for the multiple vacation model.} \end{cases}$$

- While the system is operational, there is a possibility of experiencing a catastrophic failure at any moment, resulting in the loss of all current jobs within the system. This catastrophic event is modeled as a Poisson process with an occurrence rate of γ. If such a disaster occurs, the system promptly enters a repair process in order to recover from the damage. The repair time of the servers, has an exponential distribution with rate η.

- During the repair period, the main servers are replaced by the substitute ones which serve new failed machines at a slow rate. The service times during this period are exponentially distributed with rate μ_R, where $\mu_R < \mu_B$.

- Upon arrival, if a customer finds some servers working and others free, he is served immediately. However, during vacation, regular busy, or repair periods, customers have the option to join the queue with a probability of θ_n or choose not to join (balk) with a probability of $\theta'_n = 1 - \theta_n$. Here, $0 \le \theta_{n+1} \le \theta_n \le 1$ holds for $c \le n \le N-1$ in the case of repair and regular busy periods, and $1 \le n \le N-1$ during the vacation period. It is important to note that $\theta_0 = 1, ..., \theta_{c-1} = 1$; for repair and regular busy periods, and $\theta_0 = 1$ for the vacation period. Additionally, it should be noted that $\theta_N = 0$ for both cases.
  Then, in the case of the vacation period, we can express the arrival rate as:

$$\lambda_n = \theta_n \lambda, \quad \text{for } 1 \le n \le N.$$

For the system in the down or regular operative mode, the arrival rate can be defined as:

$$\lambda_n = \begin{cases} \lambda, & \text{for } n < c, \\ \theta_n \lambda, & \text{for } c \le n \le N. \end{cases}$$

- Upon arrival, if the servers are in the regular working or working breakdown period, the customer activates an impatience timer, either T_B or T_R depending on the period. If the customer's service is not completed before the timer expires, the customer has the option to abandon the system. During the vacation period, a new arrival activates its own timer, denoted as T_V. Customers may choose to give up if the service is unavailable before their impatience timer expires. The impatience times, denoted as T_V, T_B, and T_R are random variables that follow exponential distributions with rates $\chi_V, \chi_B, \chi_R > 0$, respectively. The customers' timers are i.i.d. random variables and independent of the number of waiting customers.

- Different stochastic processes involved in the system are supposed to be independent of each other.

3.3 Motivation and illustration of the queueing model for an AMS

In this section, we motivate and illustrate the practical application of the proposed queueing model by considering an Automatic Manufacturing System (AMS) as an example. An AMS consists of several machines or workstations that process the arrival of jobs or products according to FCFS policy. These machines can be modeled as servers in the queueing system, and the jobs or products can be modeled as customers. The AMS is subject to various modes and customer behaviors that affect its performance and cost.

One of the failure modes is a disaster, which may occur suddenly in the form of an electrical failure, fire, cyber attack, or other event causing major damage to the production facility. Such a disaster results in all jobs or products waiting or currently being processed in the system being lost. This is where our queueing model comes into play. It models this disaster as a Poisson process, allowing to predict and plan for such events. Then, it is necessary to develop a repair and recovery plan that minimizes the downtime and loss of work-in-progress inventory. During the repair period following a disaster, new arrivals may be served at a lower rate (working breakdown). Our model captures this through the use of substitute servers, which serve new failed machines at a slow rate during the repair process. This ensures that production continues as smoothly as possible, maintaining the efficiency of the system and avoiding the loss of potential customers.

Another factor that affects the AMS is vacation periods, which may occur when the machines need downtime for preventive maintenance, upgrades, or periodic breaks after completing all current jobs. During vacations, the machines take a temporary leave from processing new jobs and return after a random period. A vacation can be either single or multiple. When the servers go, simultaneously, on break, they leave the system temporarily and return after a random time. Before going on vacation, the servers may wait for a while if there are no customers in the system. A vacation can have both positive and negative effects on the AMS. On one hand, it can improve the performance of the servers. On the other hand, it can increase the workload and waiting time of the remaining servers and customers. Therefore, it is essential to schedule vacations appropriately and balance their benefits and costs. Our queueing model can help with this task by providing useful performance measures such as the expected number of customers in the system during a break period, the expected waiting time of a customer during a break period, and the expected number of impatient customers during a break period.

Customer behavior also affects the AMS. Customers may choose not to enter the system or leave the system before being served due to various reasons. These behaviors are known as balking and reneging, respectively. Balking occurs when a customers decides not to join the queue because he perceives it to be too long or too slow. Reneging occurs when a customers abandons his place in the queue because he becomes impatient or dissatisfied with their waiting time. Both balking and reneging can cause financial losses for the AMS, as there may be costs associated with preparing the order or reserving the place in the queue. Moreover, balking and reneging can reduce the demand and reputation of the AMS, as well as the satisfaction and loyalty of the customers. Therefore, it is crucial to prevent or reduce balking and reneging by improving the service quality and efficiency of the AMS, as well as the satisfaction and loyalty of the customers.

Therefore, it is well noted that our model can provide valuable insights and guidance for practitioners and researchers who are interested in improving the performance and cost-effectiveness of such systems. Figure 3.1 exemplifies the automated manufacturing system.

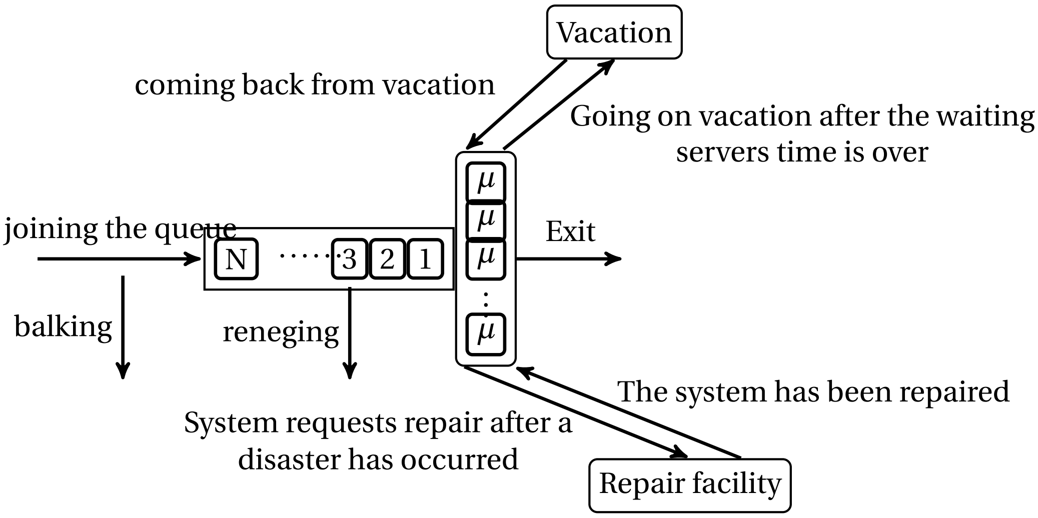

Figure 3.1: An AMS modeled as $M/M/c/N$ queue with single and multiple vacations, waiting servers, disasters, and working breakdowns

3.4 Derivation of the steady-state distribution

We formulate the process as a quasi birth-and-death(QBD) process. model at time t is defined by $\{J(t), N(t); t \geq 0\}$. Let $J(t)$ denote the server state at time t such that

$$J(t) = \begin{cases} 0, & \text{if the the system is on vacation (V);} \\ 1, & \text{if the system is in a regular busy period;} \\ 2, & \text{if the system is in a reparation period,} \end{cases}$$

and $N(t)$ is the number of customers in the system at that time. Thus $\{J(t), N(t); t \geq 0\}$ is a bivariate Markov process, the corresponding state space is

$$\Omega = \left\{ \begin{array}{ll} (j,n): & j,0,1,2,\cdots \\ & n = 0,1,\cdots,N \end{array} \right\}.$$

To study the steady-state behavior of the queueing system, let $\pi_{j,n} = \lim_{t\to\infty} \mathbb{P}\{J(t) = j; N(t) = n\}$ be the steady-state system probabilities of our system, where $(j;n) \in \Omega$. Next, we depict the transition rate diagram of proposed queueing system (see Figure 3.2). For this, the following notion is needed:

$$\psi_{0,n} = n\chi_V \ , \psi_{1,n} = \begin{cases} n(\mu_B + \chi_B), & 1 \leq n \leq c-1; \\ c\mu_B + n\chi_B, & c \leq n \leq N; \end{cases}$$

$$\psi_{2,n} = \begin{cases} n(\mu_R + \chi_R), & 1 \leq n \leq c-1; \\ c\mu_R + n\chi_R, & c \leq n \leq N. \end{cases}$$

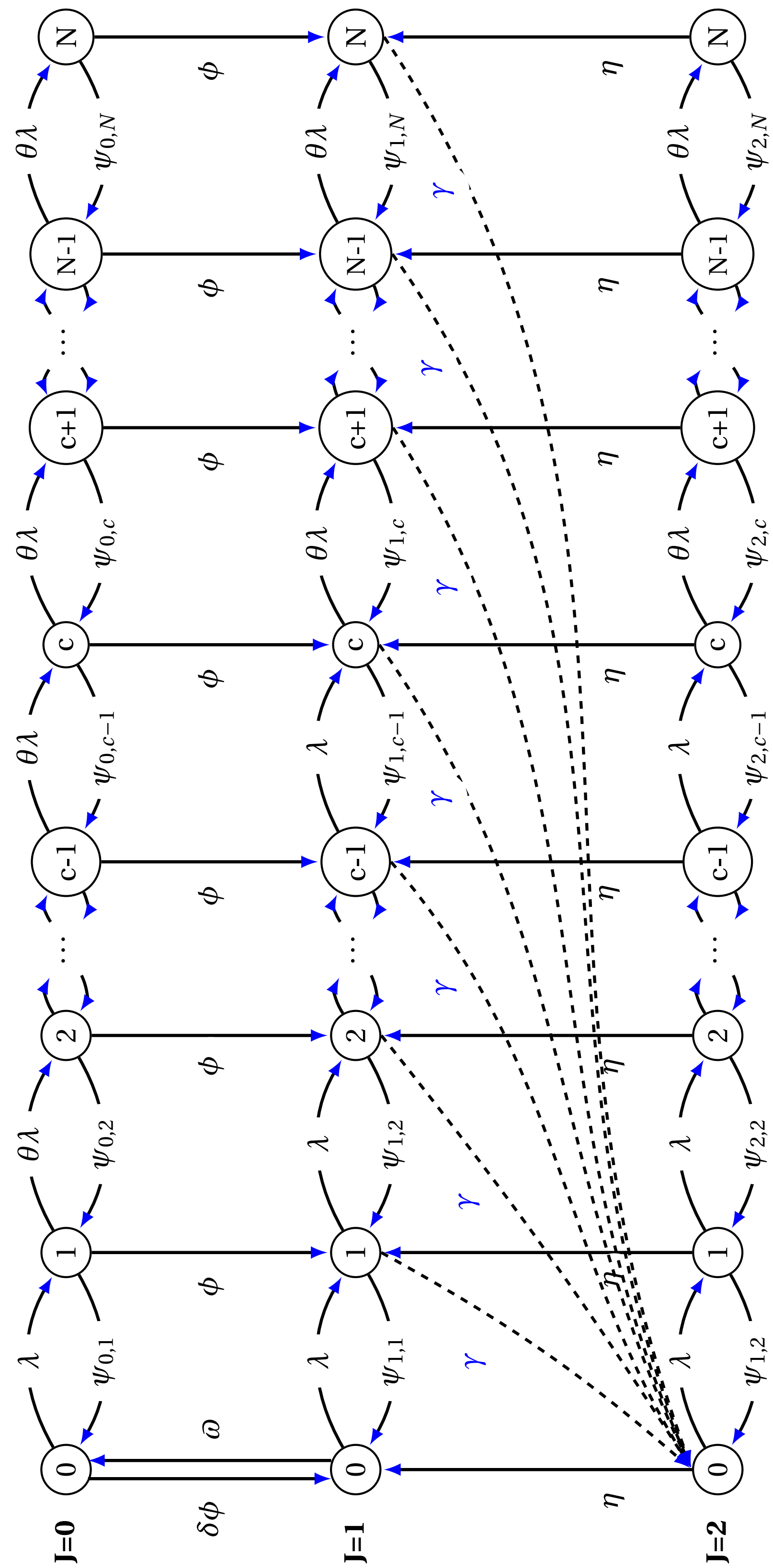

Figure 3.2: Transition diagram.

Then, we set the Chapman-Kolmogorov equations specifying the Markov model:

$$
\begin{aligned}
(\lambda+\delta\phi)\pi_{0,0} &= \varpi\pi_{1,0}+\chi_V\pi_{0,1} \quad n=0,\\
(\theta_1\lambda+\phi+\chi_V)\pi_{0,1} &= \lambda\pi_{0,0}+2\chi_V\pi_{0,2}, \quad n=1,\\
(\theta_n\lambda+\phi+n\chi_V)\pi_{0,n} &= \theta_{n-1}\lambda\pi_{n-1,0}+(n+1)\chi_V\pi_{0,n+1}, \quad 2\le n\le N-1,\\
(\phi+N\chi_V)\pi_{0,N} &= \theta_{N-1}\lambda\pi_{0,N-1}, \quad n=N,\\
(\lambda+\varpi)\pi_{1,0} &= \eta\pi_{2,0}+\chi_B\pi_{1,1}+\delta\phi\pi_{0,0}, \; n=0,\\
(\lambda+\gamma+n(\mu_B+\chi_B))\pi_{1,n} &= \lambda\pi_{1,n-1}+(n+1)(\mu_B+\chi_B)\pi_{1,n+1}+\eta\pi_{2,n}\\
&\quad +\phi\pi_{0,n}, \quad 1\le n\le c-1,\\
(\lambda\theta_c+\gamma+n(\mu_B+\chi_B))\pi_{1,n} &= \lambda\pi_{1,n-1}+(c\mu_B+(n+1)\chi_B)\pi_{1,n+1}+\eta\pi_{2,n}\\
&\quad +\phi\pi_{0,n}, \quad n=c,\\
(\lambda\theta_n+\gamma+c\mu_B+n\chi_B)\pi_{1,n} &= \theta_{n-1}\lambda\pi_{1,n-1}+(c\mu_B+(n+1)\chi_B)\pi_{1,n+1}+\eta\pi_{2,n}\\
&\quad +\phi\pi_{0,n}, \quad c+1\le n\le N-1,\\
(\gamma+c\mu_B+N\chi_B)\pi_{1,N} &= \theta_{N-1}\lambda\pi_{1,N-1}+\eta\pi_{2,N}+\phi\pi_{0,N}, \quad n=N,\\
(\lambda+\eta)\pi_{2,0} &= \gamma\sum_{n=1}^{N}\pi_{1,n}+(\mu_R+\chi_R)\pi_{2,1}, \quad n=0,\\
(\lambda+\eta+n(\mu_R+\chi_R))\pi_{2,n} &= \lambda\pi_{2,n-1}+(n+1)(\mu_R+\chi_R)\pi_{2,n+1}, \; 2\le n\le c-1,\\
(\lambda\theta_c+\eta+n(\mu_R+\chi_R))\pi_{2,n} &= \lambda\pi_{2,n-1}+(c\mu_R+(n+1)\chi_R)\pi_{2,n+1}, \; n=c,\\
(\lambda\theta_c+\eta+(c\mu_R+n\chi_R))\pi_{2,n} &= \theta_{n-1}\lambda\pi_{2,n-1}+(c\mu_R+(n+1)\chi_R)\pi_{2,n+1},\\
&\quad c+1\le n\le N-1,\\
(\eta+c\mu_R+N\chi_R))\pi_{2,N} &= \theta_{N-1}\lambda\pi_{2,N-1}, \quad n=N.
\end{aligned}
$$

The normalizing condition is as: $\sum_{n=0}^{N}\pi_{0,n}+\sum_{n=0}^{N}\pi_{1,n}+\sum_{n=0}^{N}\pi_{2,n}=1.$

The following notions are needed for the sequel of analysis:

$$
\alpha_n=\begin{cases}-(\lambda+\delta\phi), & n=0,\\ -(\theta_n\lambda+\psi_{0,n}+\phi), & 1\le n\le N-1,\\ -(\psi_{0,N}+\phi), & n=N,\end{cases}
$$

$$
\beta_n=\begin{cases}-(\lambda+\varpi), & n=0,\\ -(\lambda+\psi_{1,n}+\gamma), & 1\le n\le c-1,\\ -(\theta_c\lambda+\psi_{1,n}+\gamma), & n=c,\\ -(\theta_n\lambda+\psi_{1,n}+\gamma), & c+1\le n\le N-1,\\ -(\psi_{1,N}+\gamma), & n=N,\end{cases}
$$

and

$$
\varepsilon_n=\begin{cases}-(\lambda+\eta), & n=0,\\ -(\lambda+\psi_{2,n}+\eta), & 1\le n\le c-1,\\ -(\theta_c\lambda+\psi_{2,n}+\eta), & n=c,\\ -(\theta_n\lambda+\psi_{2,n}+\eta), & c+1\le n\le N-1,\\ -(\psi_{2,N}+\eta), & n=N.\end{cases}
$$

After having arranged the system's states, we construct the infinitesimal generator matrix denoted by Q:

$$Q = \begin{pmatrix} \Lambda_1 & \Lambda_2 & \Lambda_3 \\ \Theta_1 & \Theta_2 & \Theta_3 \\ \Omega_1 & \Omega_2 & \Omega_3 \end{pmatrix},$$

with

$$\Lambda_1 = \begin{pmatrix} \alpha_0 & \lambda_0 & & & & \\ \psi_{0,1} & \alpha_1 & \lambda_1 & & & \\ & \psi_{0,2} & \alpha_2 & \lambda_2 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \psi_{0,N-1} & \alpha_{N-1} & \lambda_{N-1} \\ & & & & \psi_{0,N} & \alpha_N \end{pmatrix}_{N+1\times N+1}, \Lambda_2 = \begin{pmatrix} \delta\phi & 0 & \dots & \dots & 0 \\ 0 & \phi & \dots & \dots & 0 \\ 0 & 0 & \dots & \dots & 0 \\ \vdots & \vdots & & & 0 \\ 0 & 0 & \dots & \dots & \phi \end{pmatrix}_{N+1\times N+1},$$

$$\Theta_1 = \begin{pmatrix} \varpi & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{pmatrix}_{N+1\times N+1}, \Theta_3 = \begin{pmatrix} 0 & 0 & \dots & \dots & 0 \\ \gamma & 0 & \dots & \dots & 0 \\ \gamma & & & & 0 \\ \vdots & & & & 0 \\ \gamma & & \dots & \dots & 0 \end{pmatrix}_{N+1\times N+1},$$

$$\Theta_2 = \begin{pmatrix} \beta_0 & \lambda & & & & & & \\ \psi_{1,1} & \beta_1 & \lambda & & & & & \\ & \psi_{1,2} & \beta_2 & \lambda & & & & \\ & & \ddots & \ddots & \ddots & & & \\ & & & \psi_{1,c-1} & \beta_{c-1} & \lambda & & \\ & & & & \psi_{1,c} & \beta_c & \lambda_c & \\ & & & & \ddots & \ddots & \ddots & \\ & & & & & \psi_{1,N-1} & \beta_{N-1} & \lambda_{N-1} \\ & & & & & & \psi_{1,N} & \beta_N \end{pmatrix}_{N+1\times N+1},$$

$$\Omega_3 = \begin{pmatrix} \varepsilon_0 & \lambda & & & & & & \\ \psi_{2,1} & \varepsilon_1 & \lambda & & & & & \\ & \psi_{2,2} & \varepsilon_2 & \lambda & & & & \\ & & \ddots & \ddots & \ddots & & & \\ & & & \psi_{2,c-1} & \varepsilon_{c-1} & \lambda & & \\ & & & & \psi_{2,c} & \varepsilon_c & \lambda_c & \\ & & & & \ddots & \ddots & \ddots & \\ & & & & & \psi_{2,N-1} & \varepsilon_{N-1} & \lambda_{N-1} \\ & & & & & & \psi_{2,N} & \varepsilon_N \end{pmatrix}_{N+1\times N+1},$$

and $\Omega_2 = \begin{pmatrix} \eta & & \\ & \ddots & \\ & & \eta \end{pmatrix}_{N+1\times N+1}.$

The matrices Λ_3 and Ω_1 are zero matrices of order $N+1 \times N+1$. Next, we concentrate on the steady-state distribution of the process $\{J(t), N(t), t \geq 0\}$. Let Π be the steady-state probability vector of the matrix Q, where $\Pi = (\Pi_0, \Pi_1, \Pi_2)$ such that $\Pi_0 = (\pi_{0,0}, \pi_{0,1}, \dots \pi_{0,N})$, $\Pi_1 = (\pi_{1,0}, \pi_{1,1}, \dots \pi_{1,N})$, $\Pi_2 = (\pi_{2,0}, \pi_{2,1}, \dots \pi_{2,N})$. The steady-state equations

$$\Pi Q = 0, \tag{3.1}$$

must be verified by the vector Π and the normalisation condition:

$$\Pi e = 1, \tag{3.2}$$

with 0 a zero row vector, $e = (e_1, e_2, e_3)$ is an $(3N+3)$ column vector with ones, and $e_j,\ j = 0,1,2$ are an (N+1) columns vectors.
Making use of Equations (3.1)-(3.2) and the fact that Λ_3 and Ω_1 are zero matrices, we obtain

$$\Pi_0\Lambda_1 + \Pi_1\Theta_1 = 0, \tag{3.3}$$
$$\Pi_0\Lambda_2 + \Pi_1\Theta_2 + \Pi_2\Omega_2 = 0, \tag{3.4}$$
$$\Pi_1\Theta_3 + \Pi_2\Omega_3 = 0, \tag{3.5}$$
$$\Pi_0 e_1 + \Pi_1 e_2 + \Pi_2 e_3 = 1. \tag{3.6}$$

The matrices Λ_2 Θ_1 and Θ_3 can be written as follows:

$$\Lambda_2 = \begin{pmatrix} \delta\phi & O_1 \\ O_2 & \phi I_N \end{pmatrix};\ \Theta_1 = \begin{pmatrix} \varpi & O_1 \\ O_2 & O_3 \end{pmatrix};\ \Theta_3 = \begin{pmatrix} O_4 & \\ \gamma J_N & O_3 \end{pmatrix},$$

where $O_i, i = 1,2,3,4$ are zero matrices and I_N an identity matrix, where O_1, O_2, O_3, O_4 of order $1\times N$, $N\times 1$, $N\times N$, $1\times N+1$ respectively, and J_N is N-dimensional column vector with ones. From equation (3.3), we get

$$\begin{aligned} \Pi_0 &= -\Pi_1 \begin{pmatrix} \varpi o \\ O_5 \end{pmatrix} \\ &= -\pi_{1,0}\varpi o, \end{aligned} \tag{3.7}$$

where O_5 is an $N\times N+1$ matrix and $o = (o_0, \tilde{o})$, such that $\tilde{o} = (o_1, ..., o_N)$ is an N row vector of the matrix Λ_1^{-1}.
From equation (3.5), we have

$$\Pi_2 = -\Pi_1\Theta_3\Omega_3^{-1}. \tag{3.8}$$

By replacing Eq. (3.7) and Eq. (3.8) into Eq. (3.4), we find $-\pi_{1,0}\varpi o\Lambda_2 + \Pi_1\tilde{\Theta} = 0$, where $\tilde{\Theta} = (\Theta_2 - \eta\Theta_3\Omega_3^{-1})$.
Therefore,

$$\begin{aligned} \Pi_1 &= \pi_{1,0}\varpi\Lambda_2 o\tilde{\Theta}^{-1} \\ &= \begin{cases} \pi_{1,0}\varpi\phi o\tilde{\Theta}^{-1}, & \delta = 1; \\ \pi_{1,0}\varpi\phi\dot{\tilde{o}}\tilde{\Theta}^{-1}, & \delta = 0, \end{cases} \end{aligned} \tag{3.9}$$

and

$$\begin{aligned} \Pi_2 &= -\Pi_1\gamma\dot{\Omega}^{-1} \\ &= \begin{cases} -\pi_{1,0}\varpi\phi o\Upsilon\tilde{\Theta}^{-1}, & \delta = 1; \\ -\pi_{1,0}\varpi\phi\dot{\tilde{o}}\Upsilon\tilde{\Theta}^{-1}, & \delta = 0. \end{cases} \end{aligned} \tag{3.10}$$

where $\dot{\tilde{o}} = (0, o_1, ..., o_N)$ and $\Upsilon = \Theta_3\Omega_3^{-1}$.
Finally, based on the above analysis, from (3.7), (3.9), and (3.10), we obtain the expressions for $\pi_{j,n}$ in terms of $\pi_{1,0}$, for single and multiple vacation policies, then use normalization condition to derive this latter. We summarize our results in the following Theorem.

Theorem 3.1. *[113] For a finite source-multiserver Markovian queueing system with single and multiple vacation policies, waiting servers, disasters, working breakdowns, and impatience, the stationary probabilities $\pi_{j,n}$ can be expressed as:*

1. *For single vacation policy* $(\delta = 1)$:

$$\pi_{j,n} = \begin{cases} -\varpi o_n \pi_{1,0}, & j = 0,\ 0 \le n \le N, \\ \varpi\phi \sum\limits_{i=0}^{N} o_i \tilde{\theta}_{in} \pi_{1,0}, & j = 1,\ 0 \le n \le N, \\ -\varpi\phi \sum\limits_{j=0}^{N} \sum\limits_{i=0}^{N} o_i \tilde{\theta}_{ij} \omega_{jn} \pi_{1,0}, & j = 2,\ 0 \le n \le N, \end{cases}$$

where

$$\pi_{1,0} = \big(-\varpi \sum_{n=0}^{N} o_n + \varpi\phi \sum_{n=0}^{N} \sum_{i=0}^{N} o_i \tilde{\theta}_{in} - \varpi\phi \sum_{n=0}^{N} \sum_{j=0}^{N} \sum_{i=0}^{N} o_i \tilde{\theta}_{ij} \omega_{jn} \big)^{-1}.$$

2. *For multiple vacation policy case* $(\delta = 0)$:

$$\pi_{j,n} = \begin{cases} -\varpi o_n \pi_{1,0}, & j = 0, 0 \le n \le N, \\ \varpi\phi \sum\limits_{0=1}^{N} \dot{\tilde{o}}_i \tilde{\theta}_{in} \pi_{1,0}, & j = 1, 0 \le n \le N, \\ -\varpi\phi \sum\limits_{j=0}^{N} \sum\limits_{i=0}^{N} \dot{\tilde{o}}_i \tilde{\theta}_{ij} \omega_{jn} \pi_{1,0}, & j = 2, 0 \le n \le N, \end{cases}$$

where

$$\pi_{1,0} = \big(-\varpi \sum_{n=0}^{N} o_n + \varpi\phi \sum_{n=1}^{N} \sum_{i=0}^{N} \dot{\tilde{o}}_i \tilde{\theta}_{in} - \varpi\phi \sum_{n=1}^{N} \sum_{j=0}^{N} \sum_{i=0}^{N} \dot{\tilde{o}}_i \tilde{\theta}_{ij} \omega_{jn} \big)^{-1},$$

such that ω_{jn} are the elements of the matrix $\Upsilon = \Theta_3\Omega_3^{-1}$ and $\tilde{\theta}_{ij}$ are the elements of matrix $\tilde{\Theta} = (\Theta_2 - \eta\Theta_3\Omega_3^{-1})$, $o = (o_0, \tilde{o})$, where $\tilde{o} = (o_1, ..., o_N)$ is an N row vector of the matrix Λ_1^{-1}, and $\dot{\tilde{o}} = (0, o_1, ..., o_N)$.

Proof. The theorem has been proved in the above. □

3.5 Performance measures

In this section, based on the steady-state probabilities, we obtain different performance measures. The results are presented as follows

- The probability that the service is available :

$$P_B = \begin{cases} \sum_{n=0}^{N} \pi_{1,n} = \varpi\phi \sum_{n=0}^{N}\sum_{i=0}^{N} o_i \tilde{\theta}_{in} \pi_{1,0}, & \delta = 1, \\ \sum_{n=0}^{N} \pi_{1,n} = \varpi\phi \sum_{n=0}^{N}\sum_{i=0}^{N} \dot{\tilde{o}}_i \tilde{\theta}_{in} \pi_{1,0}, & \delta = 0. \end{cases}$$

- The probability that the service is unavailable due to vacation :

$$P_V = \sum_{n=0}^{N} \pi_{0,n} = -\varpi \sum_{n=0}^{N} o_n \pi_{1,0}.$$

- The probability that the service is under repair :

$$\begin{aligned} P_R &= \sum_{n=0}^{N} \pi_{2,n} \\ &= \begin{cases} -\varpi\phi \left(\sum_{n=0}^{N}\sum_{j=0}^{N}\sum_{i=0}^{N} o_i \tilde{\theta}_{ij} \omega_{jn} \right) \pi_{1,0}, & \delta = 1, \\ -\varpi\phi \left(\sum_{n=0}^{N}\sum_{j=0}^{N}\sum_{i=0}^{N} \dot{\tilde{o}}_i \tilde{\theta}_{ij} \omega_{jn} \right) \pi_{1,0}, & \delta = 0. \end{cases} \end{aligned}$$

- The mean system size :

$$\begin{aligned} L_s &= \sum_{n=1}^{N} n(\pi_{0,n} + \pi_{1,n} + \pi_{2,n}) \\ &= \begin{cases} -\varpi \left(\sum_{n=1}^{N} n o_n - \phi \sum_{n=1}^{N}\sum_{i=0}^{N} n o_i \tilde{\theta}_{in} + \phi \sum_{n=1}^{N}\sum_{j=1}^{N}\sum_{i=0}^{N} n o_i \tilde{\theta}_{ij} \omega_{jn} \right) \pi_{1,0}, & \delta = 1, \\ -\varpi \left(\sum_{n=1}^{N} n o_n - \phi \sum_{n=1}^{N}\sum_{i=0}^{N} n \dot{\tilde{o}}_i \tilde{\theta}_{in} + \phi \sum_{n=1}^{N}\sum_{j=0}^{N}\sum_{i=0}^{N} n \dot{\tilde{o}}_i \tilde{\theta}_{ij} \omega_{jn} \right) \pi_{1,0}, & \delta = 0. \end{cases} \end{aligned}$$

- The effective arrival rate :

$$\dot{\lambda} = \sum_{n=0}^{N} \lambda_n (\pi_{0,n} + \pi_{1,n} + \pi_{2,n}).$$

- The mean waiting time of a customer in the system :

$$W_s = L_s / \dot{\lambda}.$$

- The average balking rate :

$$R_b = \lambda - \dot{\lambda}.$$

- The system reliability :

$$\begin{aligned} P_{re} &= 1 - P_R = 1 - \sum_{n=0}^{N} \pi_{2,n} \\ &= \begin{cases} 1 + \varpi\phi \left(\sum_{n=0}^{N}\sum_{j=0}^{N}\sum_{i=0}^{N} o_i \tilde{\theta}_{ij} \omega_{jn} \right) \pi_{1,0}, & \delta = 1, \\ 1 + \varpi\phi \left(\sum_{n=0}^{N}\sum_{j=0}^{N}\sum_{i=0}^{N} \dot{\tilde{o}}_i \tilde{\theta}_{ij} \omega_{jn} \right) \pi_{1,0}, & \delta = 0. \end{cases} \end{aligned}$$

- The average reneging rate :

$$
\begin{aligned}
R_{ren} &= \chi_V \sum_{n=1}^{N} n\pi_{0,n} + \chi_B \sum_{n=1}^{N} n\pi_{1,n} + \chi_R \sum_{n=1}^{N} n\pi_{2,n} \\
&= \begin{cases} -\varpi\left(\chi_V \sum_{n=1}^{N} no_n - \chi_B\phi \sum_{n=1}^{N}\sum_{i=0}^{N} no_i\tilde{\theta}_{in} + \chi_R\phi \sum_{n=1}^{N}\sum_{j=0}^{N}\sum_{i=0}^{N} no_i\tilde{\theta}_{ij}\omega_{jn}\right)\pi_{1,0}, & \delta = 1, \\ -\varpi\left(\chi_V \sum_{n=1}^{N} no_n - \chi_B\phi \sum_{n=1}^{N}\sum_{i=0}^{N} n\dot{\tilde{o}}_i\tilde{\theta}_{in} + \chi_R\phi \sum_{n=1}^{N}\sum_{j=0}^{N}\sum_{i=0}^{N} n\dot{\tilde{o}}_i\tilde{\theta}_{ij}\omega_{jn}\right)\pi_{1,0}, & \delta = 0. \end{cases}
\end{aligned}
$$

- The mean number of customers served per unit time :

$$
E_s = \mu_B\left(\sum_{n=1}^{c-1} n\pi_{1,n} + c\sum_{n=c}^{N} \pi_{1,n}\right) + \mu_R\left(\sum_{n=1}^{c-1} n\pi_{2,n} + c\sum_{n=c}^{N} \pi_{2,n}\right).
$$

3.6 Optimization analysis

Now, we focus on the cost parameter optimization. We aim to find the values of the decision variables (optimum system capacity, minimum number of servers, and optimum service rates during both repair and regular busy periods) that minimize the objective function (the total expected cost of the system). Combined with the above results, we then investigate the effects of main system parameters on different performance measures given above as well as on the expected cost function per unit time T_c and on the total expected profit per unit time of the system T_{ep} that are given as:

$$
\begin{aligned}
\mathrm{T}_c &= c_B P_B + c_V P_V + c_R P_R + c_h E(L_s) + c_l(R_{ren} + R_b) + c \times c_f \\
&\quad + c \times (\mu_B c_{s_1} + \mu_R c_{s_2}),
\end{aligned} \tag{3.11}
$$

where

c_B	=	Cost per unit time when the system is in normal busy period;
c_V	=	Cost per unit time when the system being in vacation period;
c_R	=	Cost per unit time when the system is under repair;
c_h	=	Holding cost per unit time when a customer enters the queue, and wait for service;
c_{s_1}	=	Cost per service per unit time in normal busy period;
c_{s_2}	=	Cost per service per unit time in repair period;
c_l	=	Cost per service per unit time when a customer is lost;
c_f	=	Fixed purchase cost of the server per unit,

and

$$
\mathrm{T}_{ep} = \mathrm{T}_{ev} - \mathrm{T}_c,
$$

$\mathrm{T}_{ev} = \mathrm{R} \times E_s$ denotes the total expected revenue per unit time of the system and R represents the revenue earned by providing service to a customer. The objective function (11) is complex and highly non-linear. In this case, it is beneficial to use the direct search method and quasi-Newton method for finding good approximations to the optimal solution. For the numerical analysis, we consider the different cost elements as: $c_V = \$18$, $c_B = \$50$, $c_R = \$60$, $c_h = \$18$, $c_{s1} = \$1$, $c_{s2} = \$1$, $c_l = \$18$, $c_f = \$1$, and $R = 50$.

3.6.1 Direct search method

In this subsection, we aim to obtain the joint optimal values (c^*, N^*) by employing the direct search method. If the function $T_c(c,N)$ is convex (unimodal), a single relative minimum exists. In order to get $T_c(c^*,N^*)$, we have to prove that the $\mathrm{T}_c(c,N)$ is convex. We take the different parameters as follows: $\lambda = 3$; $\chi_V = 0.5$; $\chi_B = 0.3$; $\chi_R = 0.7$; $\mu_R = 1$; $\mu_B = 4$; $\phi = 1.2$; $\eta = 0.5$; $\gamma = 0.2$; $\varpi = 0.5$; $\theta = 1 - \frac{n}{N}$ and vary the system capacity (N) as well as the number of servers in the system (c). The obtained numerical results have been presented in the following fiqures and Tables.

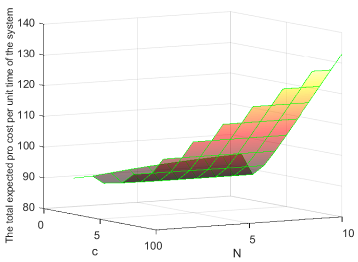

Figure 3.3: The expected cost vs. c & N for SVP

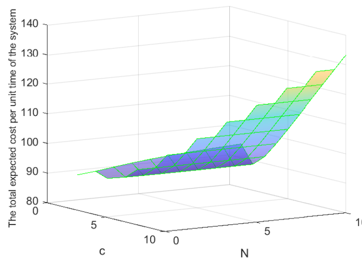

Figure 3.4: The expected cost vs. c &N for MVP

Table 3.1: c & N for various T_c

SVP	N/c	1	2	3	4	5	6	7	8
	1	90.1648	-	-	-	-	-	-	-
	2	90.4144	88.7606	-	-	-	-	-	-
	3	91.0930	88.4810	91.9394	-	-	-	-	-
	4	91.7545	88.3668	91.9066	97.1769	-	-	-	-
	5	92.3352	88.3182	91.8922	97.2018	103.0250	-	-	-
	6	92.8344	88.2993	91.8862	97.2214	103.0509	109.0118	-	-
	7	93.2633	**88.2949**	91.8845	97.2372	103.0712	109.0322	115.0239	-
	8	93.6337	88.2979	91.8849	97.2501	103.0874	109.0484	115.0399	121.0381
MVP	1	89.7226	-	-	-	-	-	-	-
	2	90.3297	88.8882	-	-	-	-	-	-
	3	91.1546	88.6897	92.1756	-	-	-	-	-
	4	91.9022	88.6248	92.1708	97.4343	-	-	-	-
	5	92.5414	**88.6103**	92.1756	97.4727	103.2904	-	-	-
	6	93.0836	88.6167	92.1837	97.5023	103.3246	109.2835	-	-
	7	93.5456	88.6320	92.1927	97.5258	103.3513	109.3097	115.3009	-
	8	93.9423	88.6508	92.2017	97.5449	103.3726	109.3306	115.3213	121.3195

Figures 3.3 and 3.4 clearly show the convexity of the curves for both single and multiple vacation policies; there exist certain values of the number of servers (c) and the system capacity (N) that minimize the total expected cost function for the chosen set of model parameters. More precisely,from Table 3.1, we have the combined optimal solution is $(c^*, N^*) = (2,7)$ with $\mathrm{T}_c = \$88.2949$, in the single vacation policy, and $(c^*, N^*) = (2,5)$ with $\mathrm{T}_c = \$88.6103$, in multiple vacation policy.

3.6.2 Quasi-Newton method

After determining the optimal values of c and N using the direct search method, we proceed to search for the optimal service rates (μ_B^*,μ_R^*). This is achieved using the quasi-Newton method. Our focus then shifts to examining the effects of different system parameters on optimization results. Next, we conduct a sensitivity analysis. This analysis evaluates the impacts of various system parameters. These parameters include the arrival rate (λ), vacation (ϕ), waiting servers rate (ϖ), disaster rate (γ), and repair rate (η). We also consider the reneging rates during vacation, busy, and repair periods, represented by ξ_V, ξ_B, and ξ_R respectively. Additionally, we look at the non-balking probability (θ_n).

The impacts of these parameters on different performance measures and $\mathrm{T}_c(c^*, N^*, \mu_B^*, \mu_R^*)$ are considered. The numerical results of this analysis are provided in Tables 3.2-3.5. Furthermore, the impact of these system parameters $\mathrm{T}_{ep}(c^*, N^*, \mu_B^*, \mu_R^*)$ is illustrated in Figures 3.5-3.10. The numerical study is presented for both SVP and MVP.

Table 3.2: ϕ and λ vs. different performance measures and T_c for $\chi_V = 0.5, \chi_B = 0.3, \chi_R = 0.7$, $\eta = 0.5, \gamma = 0.2, \varpi = 0.5$

	ϕ	λ	c^*	N^*	μ_B^*	μ_R^*	P_V	P_B	P_R	L_s	W_s	R_b	R_{ren}	P_{re}	E_s	T_c
SVP	1	3.5	2	4	6.5707	1.3195	0.1962	0.6862	0.1175	0.8085	0.2677	0.4804	0.3550	0.8824	4.0728	92.0711
		4	2	4	6.9933	1.6311	0.1894	0.6872	0.1234	0.8566	0.2511	0.5883	0.3760	0.8766	4.6001	97.2001
		4.5	2	4	7.4020	1.9396	0.1834	0.6881	0.1285	0.8987	0.2366	0.7009	0.3942	0.8715	5.1145	101.9905
	1.5	3.5	2	10	6.9049	1.6298	0.1425	0.7346	0.1229	0.7833	0.2360	0.1813	0.3385	0.8771	4.5158	89.1947
		4	2	7	7.4295	1.8682	0.1386	0.7341	0.1273	0.8058	0.2180	0.3038	0.3484	0.8727	5.0619	93.6769
		4.5	2	6	7.9318	2.1329	0.1350	0.7336	0.1315	0.8325	0.2037	0.4134	0.3600	0.8685	5.6114	98.0306
	2	3.5	2	13	7.1081	1.6889	0.1132	0.7628	0.1240	0.7168	0.2121	0.1202	0.3060	0.8760	4.6930	87.7859
		4	2	9	7.6670	1.9484	0.1100	0.7614	0.1286	0.7468	0.1969	0.2081	0.3192	0.8714	5.2793	91.9320
		4.5	2	7	8.2296	2.1929	0.1074	0.7603	0.1323	0.7686	0.1835	0.3105	0.3288	0.8677	5.8540	96.0738
MVP	1	3.5	2	4	6.3678	1.2905	0.2386	0.6457	0.1157	0.8699	0.2940	0.5413	0.3856	0.8843	3.8935	93.1785
		4	2	4	6.7694	1.6092	0.2248	0.6532	0.1221	0.9141	0.2731	0.6527	0.4043	0.8779	4.4117	98.2658
		4.5	2	4	7.1620	1.9237	0.2134	0.6589	0.1277	0.9527	0.2553	0.7683	0.4205	0.8723	4.9235	103.1658
	1.5	3.5	2	8	6.8076	1.5410	0.1923	0.6880	0.1197	0.8361	0.2576	0.2539	0.3655	0.8803	4.3304	89.9404
		4	2	7	7.2652	1.8421	0.1800	0.6943	0.1256	0.8726	0.2389	0.3468	0.3809	0.8744	4.8878	94.5163
		4.5	2	6	7.7347	2.1122	0.1707	0.6991	0.1302	0.8943	0.2217	0.4653	0.3900	0.8698	5.4270	99.0260
	2	3.5	2	12	7.0610	1.6350	0.1672	0.7118	0.1210	0.7798	0.2329	0.1518	0.3372	0.8790	4.5440	88.0928
		4	2	8	7.5837	1.8809	0.1563	0.7177	0.1260	0.7996	0.2141	0.2652	0.3456	0.8740	5.1072	92.5729
		4.5	2	7	8.0844	2.1661	0.1471	0.7222	0.1306	0.8276	0.1996	0.3538	0.3575	0.8694	5.6821	96.7991

Table 3.3: γ, ϖ, and η vs. different performance measures and T_c for $\chi_V = 0.5, \chi_B = 0.3, \chi_R = 0.7$, $\lambda = 3, \phi = 1.2$

	γ	ϖ	η	c^*	N^*	μ_B^*	μ_R^*	P_V	P_B	P_R	L_s	W_s	R_b	R_{ren}	P_{re}	E_s	T_c
SVP	0.2	0.6	0.8	2	17	6.2955	0.2096	0.2093	0.7163	0.0744	0.8803	0.3050	0.1142	0.3806	0.9256	3.7719	83.8095
			1	2	35	6.1974	0.0001	0.2101	0.7283	0.0616	0.9063	0.3080	0.0577	0.3842	0.9384	3.8184	82.5542
		0.8	0.8	2	10	6.2265	0.0834	0.2557	0.6721	0.0721	0.9083	0.3250	0.2053	0.3969	0.9279	3.6018	84.3462
			1	2	15	6.0834	0.0001	0.2559	0.6841	0.0600	0.9350	0.3271	0.1419	0.3999	0.9400	3.6523	83.1603
	0.4	0.6	0.8	2	12	5.9993	1.3056	0.1930	0.6667	0.1404	0.8398	0.2945	0.1481	0.3718	0.8596	3.5953	86.3136
			1	2	44	6.0669	0.8200	0.1974	0.6861	0.1165	0.9080	0.3073	0.0456	0.3978	0.8835	3.6570	84.9475
		0.8	0.8	2	8	5.9284	1.2086	0.2364	0.6273	0.1363	0.8655	0.3130	0.2347	0.3863	0.8637	3.4395	86.8294
			1	2	16	6.0019	0.7196	0.2416	0.6453	0.1131	0.9354	0.3262	0.1320	0.4135	0.8869	3.5003	85.5008
MVP	0.2	0.6	0.8	2	9	6.1993	0.0462	0.2718	0.6568	0.0714	0.9201	0.3325	0.2330	0.4033	0.9286	3.5451	84.5228
			1	2	13	6.0446	0.0001	0.2715	0.6690	0.0595	0.9468	0.3342	0.1671	0.4061	0.9405	3.5978	83.3578
		0.8	0.8	2	6	6.0559	0.0001	0.3242	0.6069	0.0689	0.9316	0.3528	0.3592	0.4110	0.9311	3.3392	85.0607
			1	2	8	5.9007	0.0001	0.3233	0.6191	0.0576	0.9653	0.3551	0.2814	0.4174	0.9424	3.3996	83.9861
	0.4	0.6	0.8	2	7	5.9024	1.1697	0.2519	0.6133	0.1348	0.8707	0.3191	0.2715	0.3896	0.8652	3.3808	87.0042
			1	2	13	5.9774	0.6843	0.2568	0.6312	0.1119	0.9426	0.3325	0.1648	0.4179	0.8881	3.4448	85.6799
		0.8	0.8	2	5	5.8095	1.0461	0.3023	0.5679	0.1298	0.8845	0.3394	0.3936	0.3989	0.8702	3.1894	87.5230
			1	2	8	5.8880	0.5675	0.3074	0.5846	0.1079	0.9623	0.3536	0.2786	0.4304	0.8921	3.2578	86.2343

Table 3.4: χ_R, χ_B, and χ_V vs. different performance measures and T_c for $\lambda = 3, \eta = 0.5, \gamma = 0.2$, $\phi = 1.2$, $\varpi = 0.5$

	χ_R	χ_B	χ_V	c^*	N^*	μ_B^*	μ_R^*	P_V	P_B	P_R	L_s	W_s	R_b	R_{ren}	P_{re}	E_s	T_c
SVP	0.1	0.4	0.3	2	3	6.3288	1.2313	0.1810	0.7091	0.1098	0.6926	0.2707	0.4420	0.2152	0.8902	3.6733	86.7194
			0.5	2	3	6.3080	1.2289	0.1813	0.7091	0.1096	0.6747	0.2619	0.4238	0.2499	0.8904	3.6547	86.6412
		0.6	0.3	2	3	6.4353	1.1848	0.1847	0.7090	0.1063	0.6815	0.2659	0.4375	0.2766	0.8937	3.6057	87.5144
			0.5	2	3	6.4134	1.1824	0.1849	0.7090	0.1061	0.6633	0.2570	0.4191	0.3118	0.8939	3.5869	87.4303
	0.3	0.4	0.3	2	3	6.3817	1.1099	0.1817	0.7091	0.1092	0.6857	0.2674	0.4355	0.2396	0.8908	3.6576	86.7539
			0.5	2	4	6.2940	1.2454	0.1801	0.7091	0.1108	0.7071	0.2661	0.3428	0.2913	0.8892	3.6831	86.5679
		0.6	0.3	2	3	6.4879	1.0613	0.1853	0.7090	0.1057	0.6749	0.2628	0.4312	0.3003	0.8943	3.5904	87.5400
			0.5	2	3	6.4658	1.0588	0.1856	0.7090	0.1055	0.6567	0.2538	0.4128	0.3355	0.8945	3.5715	87.4553
MVP	0.1	0.4	0.3	2	3	6.1713	1.1966	0.2346	0.6581	0.1073	0.7518	0.3020	0.5107	0.2329	0.8927	3.4993	87.2180
			0.5	2	3	6.1319	1.1905	0.2390	0.6542	0.1068	0.7327	0.2921	0.4920	0.2801	0.8932	3.4617	87.1493
		0.6	0.3	2	2	6.3540	0.8976	0.2443	0.6550	0.1007	0.6651	0.2823	0.6439	0.2667	0.8993	3.3380	88.0555
			0.5	2	2	6.3175	0.8894	0.2496	0.6502	0.1002	0.6525	0.2749	0.6264	0.3092	0.8998	3.3012	88.0154
	0.3	0.4	0.3	2	3	6.2240	1.0731	0.2355	0.6579	0.1066	0.7452	0.2986	0.5045	0.2571	0.8934	3.4833	87.2463
			0.5	2	3	6.1845	1.0666	0.2400	0.6539	0.1061	0.7261	0.2888	0.4858	0.3044	0.8939	3.4456	87.1763
		0.6	0.3	2	2	6.3994	0.7920	0.2450	0.6548	0.1002	0.6601	0.2793	0.6366	0.2882	0.8998	3.3250	88.0741
			0.5	2	3	6.2854	1.0167	0.2446	0.6527	0.1026	0.7160	0.2844	0.4822	0.3650	0.8974	3.3773	87.9385

Table 3.5: θ_n vs. different performance measures, T_c, and T_p for $\lambda = 3, \eta = 0.5, \gamma = 0.2, \phi = 1.2, \varpi = 0.5$, $\chi_V = 0.5, \chi_B = 0.3, \chi_R = 0.7$

	θ_n	c^*	N^*	μ_B^*	μ_R^*	P_V	P_B	P_R	L_s	W_s	R_b	R_{ren}	P_{re}	E_s	T_c	T_{ep}
SVP	$\frac{1}{n+1}$	1	1	9.1546	0.0001	0.2187	0.7080	0.0733	0.3751	0.1539	0.5626	0.1614	0.9267	6.4818	73.6724	250.4176
	$1-\frac{n}{N}$	2	7	6.2488	1.1907	0.1766	0.7092	0.1142	0.7623	0.2741	0.2189	0.3325	0.8858	3.7517	86.0172	101.5661
	$1-(\frac{n}{N})^2$	2	5	6.2026	1.2486	0.1752	0.7093	0.1156	0.7862	0.2779	0.1706	0.3431	0.8844	3.7782	85.8523	103.0591
MVP	$\frac{1}{n+1}$	1	1	8.8091	0.0001	0.2984	0.6317	0.0698	0.4149	0.1745	0.6224	0.1825	0.9302	5.5651	72.9149	205.3381
	$1-\frac{n}{N}$	2	5	6.1487	1.0532	0.2337	0.6563	0.1100	0.7957	0.2983	0.3325	0.3512	0.8900	3.5396	86.6557	90.3234
	$1-(\frac{n}{N})^2$	2	4	6.1121	1.1163	0.2309	0.6576	0.1115	0.8255	0.3026	0.2718	0.3646	0.8885	3.5775	86.4957	92.3810

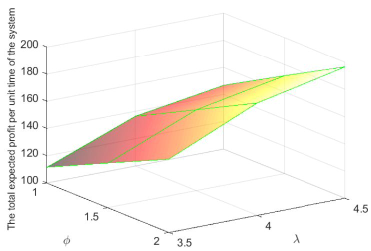

Figure 3.5: T_{ep} near $(c^*, N^*, \mu_B^*, \mu_R^*)$ for $\delta = 1$ and different values of ϕ & λ

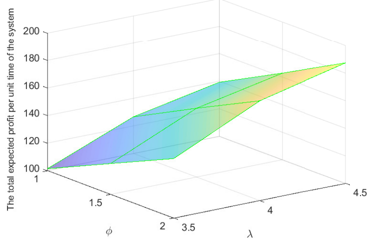

Figure 3.6: T_{ep} near $(c^*, N^*, \mu_B^*, \mu_R^*)$ for $\delta = 0$ and different values of ϕ & λ

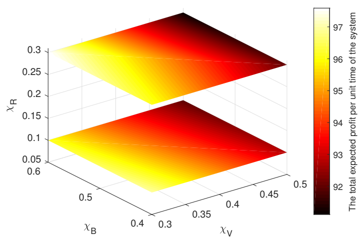

Figure 3.9: T_{ep} near $(c^*, N^*, \mu_B^*, \mu_R^*)$ for different values of χ_V, χ_B & χ_R and $\delta = 1$

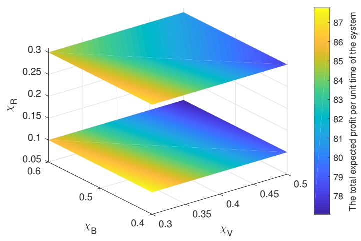

Figure 3.10: T_{ep} near $(c^*, N^*, \mu_B^*, \mu_R^*)$ for different values of χ_V, χ_B & χ_R and $\delta = 0$

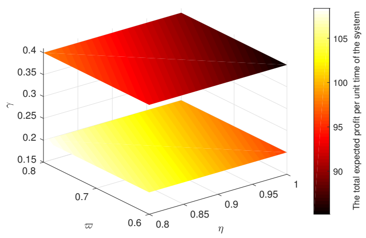

Figure 3.7: T_{ep} near $(c^*, N^*, \mu_B^*, \mu_R^*)$ for different values of η, ϖ & γ and $\delta = 1$

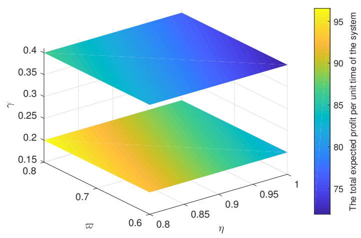

Figure 3.8: T_{ep} near $(c^*, N^*, \mu_B^*, \mu_R^*)$ for different values of η, ϖ & γ and $\delta = 0$

3.6.2.1 Discussion

The cost optimization analysis of the queueing model can help determine optimal values for the capacity, number of servers, and service rates in the AMS to minimize overall costs. The optimized system capacity N^* guides setting the number of machines/ workstations in the production line to balance investment and inventory holding costs. The optimal number of servers c^* helps determine the server capacity that minimizes costs due to disruptions and customer abandonment. The optimized service rates μ_B, μ_R help define scheduling policies for production and repair operations that maximize throughput while controlling labor expenses.

Taking into account various cost tradeoffs, the optimization framework identifies the ideal combination of capacity, staffing, and service rates for AMS to maintain high quality, throughput, and customer satisfaction while keeping costs in control. The visualization of costs further helps illustrate the impact of different decisions and failure dynamics on the bottom line. This connects the analytical model to tangible planning, configuration, and control policies for real-world flexible manufacturing systems. From the above results, Tables 3.2-3.5 and Figures 3.5-3.10, the following impacts are observed:

- The total cost and the total profit increase whenever λ increases. Since the arrival rate is growing up, there will be a high number of customers arriving in the system ($L_s \nearrow$). Thus, the total cost will be increased along with the total profit. This is indeed the case as the higher number of jobs entering the service unit will be placed and served resulting in the enhanced total profit.

- (If the vacation completion rate (ϕ) increases, the respective total expected cost (total expected profit) decreases (increases). This is because an increase in the vacation completion rate leads to a decrease in the server's vacation time ($P_V \searrow$), and thus an increase in its availability to serve customers ($P_B \nearrow$). Therefore, ($E_s \nearrow$). This in turn reduces the number of customers in the system ($L_s \searrow$), resulting in a decrease in Tc and in an increase in T_p. This is obvious, as the machines are available for production and processing, the queueing time will be less and further generated more profits from the system point of view.

- The waiting servers rate (ϖ) impacts the total cost by raising it as it increases: A decrease in the waiting server rate can increase the waiting time for customers ($W_s \nearrow$), which can augment customer impatience and a decreases in the product value. This in turn increases the total expected cost and decreases the total expected profit.

- The increase of the failure rate (γ) means that fewer customers are being served, which can lead to a decrease in the mean number of customers served ($E_s \searrow$). This leads to a decrease in revenue and an increase in costs and a decrease in the total expected profit.

- Obviously, high repair rate (η) has a nice impact on the total expected cost function and total expected profit. When the repair rate increases, it means that the failure can be quickly addressed and resolved. This leads to less downtime which results in a reduction of the repair time ($P_R \searrow$) and an increase in the station's availability to serve customers ($P_B \nearrow$). As a result, the system can serve more customers and generate more revenue, which can increase the total expected profit. Additionally, a higher repair rate can lead to lower maintenance costs, as regular maintenance can prevent failures and reduce the need for repairs. This can also contribute to a decrease in the total expected cost function. However, it is important to note that an excessively high repair rate can also lead to higher costs. This is because maintaining a high repair rate may require expensive equipment or highly skilled technicians. Moreover, there may be diminishing returns as the cost of increasing the repair rate further may outweigh the benefits. Therefore, it is important to balance the repair rate with the costs involved in order to optimize the total expected cost function and total expected profit.

- The impatience rates of customers (χ_V, χ_B, χ_R) have several negative impacts on the system's performance and profitability; abandoning the system can result in a decrease in the total number of customers served and revenue generated, which can negatively impact the system's profitability.

- A higher non-balking probability (θ_n) indicates a higher willingness of customers to join the system, which can increase the number of customers served and the revenue generated. This can positively impact the optimal performance measures and increase the total expected profit.

- From the above observations it is important to note that the majority of the results agree with our intuition while others are not so straightforward to interpret. For instance, we remark that when the joining probability increases from $\theta_n = \frac{1}{n+1}$ to $\theta_n = 1 - \frac{n}{N}$, the total expected profit decreases. This contradicts our intuition, and it can be due to the chosen costs and parameters. It is well known that higher join probability would normally increase profitability (more customers being served). But in this case, perhaps the higher loads lead to disproportionately higher holding costs that outweigh the gains.

- Further, the values of P_V , R_{ren}, R_b, L_s, W_s, P_{re}, T_c for single vacation are lower than those for multiple vacation, and the values of P_R, P_B, E_s, T_{ep} for multiple vacation are lower than those for single vacation.

Remark

The obtained results align with intuitive expectations, the single vacation model demonstrates better performance measures versus the multiple vacation model.

3.7 Conclusion

In this chapter, we proposed a queueing model tailored to machining systems within automated manufacturing environments. Our model integrates critical factors, including disasters, working breakdowns, single and multiple vacations, waiting servers, and customer impatience (balking and reneging). Unlike existing literature models that often isolate specific dynamics, our unified framework captures the interplay of multiple uncertainties. Our contributions extend beyond theory. By directly applying our queueing model to automated manufacturing systems (AMS), production and quality managers gain practical insights. They can optimize critical parameters such as system capacity, the number of servers, and service rates during disruptions to maximize throughput while balancing costs. Moreover, our customer-centric design accounts for behaviors like impatience and service refusal, ensuring AMS facilities meet operational goals while enhancing overall effectiveness.

4 Performance analysis of power saving mechanisms in mobile wireless communication: A queueing model with vacations, impatience, and disruptions

4.1 Introduction

In an era of exponential growth in mobile data consumption, the telecommunications industry faces a critical challenge: balancing the demand for high-performance networks with the imperative of energy efficiency. The Information and Communication Technology (ICT) sector, which already consumes approximately 4% of the global annual energy production [81], is poised for even greater power demands, particularly in mobile networks [33]. This surge not only impacts the battery life of mobile devices but also significantly inflates operational costs for service providers. As we stand at this crucial juncture, our research addresses a pressing need: developing sophisticated, realistic models for power-saving mechanisms in modern wireless networks.

Our study focuses on the IEEE 802.16e standard's Power Saving Mechanism, a crucial feature in wireless networks aimed at improving energy efficiency while maintaining communication reliability. It is worth pointing out that the present chapter extends the work of [98], where the authors studied customer impatience in an M/M/1 queueing system with waiting server and differentiated vacations. In contrast, this chapter considers a more complex model with state-dependent balking, catastrophes, repairs, and continued service during breakdowns. While the prior work focused on transient conditions, our analysis examines the stability and steady-state performance of this extended system, providing a broader perspective on the model's practical applicability. (A detailed description of this mechanism is provided in Section 4.2.) We propose a groundbreaking approach by modeling this mechanism through the lens of queueing theory, specifically as an $M/M/1$ queueing system. This model incorporates several features:

1. A differentiated vacation (DV) policy with waiting servers, mirroring the multi-core architecture of modern devices and the power-saving classes defined in the IEEE 802.16e standard.

2. Impatient customers, reflecting real-world packet loss due to buffer overflows and delays, implemented through state-dependent balking and reneging mechanisms.

3. Catastrophes, accounting for unpredictable network disruptions such as virus attacks, natural disasters, or hardware failures.

4. Working breakdowns, representing scenarios where the system continues to function but at a diminished capacity, simulating partial network failures or degraded performance due to interference or minor hardware issues.

The uniqueness of our approach lies in its comprehensive integration of these elements, providing a more realistic representation of network dynamics. Our model not only captures the essence of the IEEE 802.16e Power Saving Mechanism but also extends it to account for real-world complexities often overlooked in traditional models. Key innovations in our model include:

- A dual-server approach with a primary server that can enter "vacation" modes (representing power-saving states) and a secondary waiting server that remains operational, ensuring a balance between energy efficiency and service quality.

- The incorporation of two distinct vacation types (Type-I and Type-II), mirroring the power-saving classes defined in the IEEE 802.16e standard.

- A detailed representation of system behavior during catastrophes, including the use of backup routers with reduced capacity and the modeling of system recovery processes.

- The modeling of working breakdowns, allowing for a more nuanced representation of system performance under various levels of operational efficiency.

- The inclusion of customer impatience through state-depending balking and reneging, reflecting the behavior of data packets in congested or delayed network conditions.

By employing advanced mathematical techniques, including matrix geometric methods, we provide a comprehensive analysis of system stability and derive steady-state solutions. Our approach yields crucial performance measures and facilitates in-depth performance evaluation, culminating in the development of a cost model with detailed analysis.

The significance of our work lies in its potential to revolutionize the design of power-saving strategies in mobile networks. By offering a nuanced understanding of system behavior under various conditions, our model paves the way for more resilient and adaptive network designs. It bridges the gap between theoretical models and real-world network behavior.

4.2 IEEE 802.16e Power Saving Mechanism overview

The IEEE 802.16e Power Saving Mechanism is a crucial feature in wireless networks, aiming to enhance energy efficiency while maintaining communication reliability. Modeled as an M/M/1/DV queueing system, it incorporates waiting servers, balking, reneging, and factors like catastrophes, repairs, and working breakdowns. In what follows we delve into its key components:

1. Power Saving Classes:

    - The IEEE 802.16e standard defines two power saving classes:

        – Power Saving Class of Type-I: In this mode, the sleep window size for the Subscriber Station (SS) increases exponentially when there is no data traffic between the SS and the Base Station (BS).

        – Power Saving Class of Type-II: Here, the sleep window size is modeled as a random variable with a constant mean, approximating the fixed-length sleep windows used for periodic traffic patterns in the actual system.

        - The overarching goal is to reduce SS power consumption while ensuring the longevity of the battery.

2. Sleep mode operation:

    - During active periods (busy mode), the SS actively communicates with the BS to send and receive data packets.

    - When no data traffic exists, the SS enters an idle state for a random duration, modeling the waiting server in our queueing system.

    - Upon completing the idle state, the SS sends a mobile sleep request (MOB-SLP-REQ) to the BS, seeking permission to enter sleep mode.

    - The BS responds with a mobile sleep response (MOB-SLP-RES), allowing the SS to transition into sleep mode.

3. BS behavior:

    - If granted permission, the BS enters Type-I sleep mode (modeled as an exponentially distributed vacation period).

- If no data traffic persists after returning from Type-I sleep mode, the BS can enter Type-II sleep mode (also modeled as an exponentially distributed vacation period).
- After completing Type-II sleep mode, it returns to normal operation if data traffic resumes; otherwise, it re-enters another Type-II and follows this procedure until at least one data packets arrives.

4. Catastrophes and repairs:

   - The system may experience breakdowns, representing sudden disruptions within the network. These could be due to hardware failures, virus attacks, or other unforeseen events.
   - When a breakdown occurs, all data packets in the system are dropped, and the system enters a repair state.
   - During repair periods, the network operates at a reduced capacity, modeled by a slower service rate.

5. Balking:

   - Balking occurs when data packets arriving at the network decide not to enter due to unfavorable conditions.
   - The probability of balking depends on the current state of the system (busy period, sleep mode, or repair period).
   - This behavior models the decision-making process of data sources when encountering potential congestion or delays.

6. Reneging:

   - Data packets may renege (leave the system) during sleep modes or repair periods if they are not served within a certain time.
   - This behavior models the impatience of data packets in the system, particularly during periods of reduced service capacity.

4.3 Literature review

Vacation queues have long been a focal point in queueing theory research, owing to their widespread applications in manufacturing, telecommunications, postal services, and file transfer systems. For comprehensive overviews, readers may consult the seminal works of [68], [36], [100], and [101]. Recent advancements in this field are explored in studies by [40], [73], [2], and [74].

A notable innovation in this domain is the differentiated vacation (DV) approach, introduced by [47] and [48]. This model distinguishes between two vacation types: a longer period initiated after serving at least one customer (type-I), and a shorter interval taken when the server returns to find an empty system (type-II). The DV concept has gained traction due to its relevance in diverse scenarios, including call centers, energy-efficient mobile technologies, hospital emergency operations, and service stations. The practical significance of DV queues has spurred further research, as evidenced in works by [18], [98], [41], [105], and [21]. These studies expand on the foundational concepts, exploring various applications and theoretical extensions of the DV model.

Catastrophic events are incorporated to represent major system failures that clear the entire queue and potentially render the server inoperable. This could model severe network outages or hardware malfunctions. Queueing models incorporating catastrophes have gained significant attention due to their relevance in various practical applications. These models find utility in diverse fields such as computer networks [25], muscle contraction processes [34], insect populations [50], broadband communication networks [51], emergency supply chain systems [45], systems with uncertainty [52], and communication and cellular networks [83]. The study of these models has intensified over the past two decades, driven by the rapid evolution of communication systems and networks. In these contexts, catastrophes often manifest as events that force the departure of all customers, including those currently being served. For instance, in computer networks, a virus can act as a delete operation, erasing all stored data. Significant contributions to this field include the analysis of $M/M/1$ queues with catastrophes [63], multi-server retrial queues with negative customers and disasters [93], and finite-source discrete-time $Geo/Geo/1$ queues with disasters [56]. Further developments encompassed $GI/Geo/1$ queues [79], $M/G/1$ queues with disasters and working breakdowns [60], and the exploration of $M/G/1$ queue and $GI/M/1$ queue in multi-phase random environments [54, 55]. Recent work has also examined discrete-time $Geo/Geo/1$ queues with feedback, repair, and disaster [95].

Customer behavior in the queueing system is modeled through the inclusion of balking and reneging mechanisms. Balking represents the scenario where incoming requests opt not to join the queue upon arrival, typically due to perceived long wait times or system congestion. This could correspond to packets being dropped immediately if the network appears overloaded. Reneging, conversely, occurs when requests already in the queue depart before receiving service, possibly due to timeout mechanisms or quality of service requirements. These features allow the model to capture more realistic network dynamics, especially under high-load conditions where service delays can significantly impact performance.

Queueing systems incorporating customer impatience have attracted considerable research interest. These models offer practical applications across various service systems and e-commerce platforms. The literature on this topic has evolved significantly, with notable contributions spanning nearly two decades [66, 69, 114, 115]. More recent studies by [13–15, 20, 107] have expanded the field, providing a comprehensive exploration of impatient customer behavior in queueing systems.

The chapter is organized as follows: The Section 4.4 is dedicated for the description of the model. The model analysis is presented in Section 4.5 at which the stability condition of the system is derived along with the system's stationary probability distribution. The performance measures are obtained in Section 4.6. Section 4.7 presents the numerical analysis. Section 4.8 concludes the chapter.

4.4 Model description

Our model represents the IEEE 802.16e Power Saving Mechanism as an $M/M/1/DV$ queueing system with additional features to capture real-world network dynamics. The system incorporates waiting servers, impatience (balking and reneging), catastrophes, repairs, and working breakdowns. Below, we provide a comprehensive description of the model components and their interactions.

4.4.1 Basic queueing structure

The core of our model is an $M/M/1/DV$ queue, where:

- Customers (data packets) follow a Poisson process to enter the system with an arrival rate of λ.
- Customers are served on a FCFS basis. The service time depends on whether the system is in busy period or repair period. During busy periods, the service time follows an exponential distribution with a rate of μ_B. While, during repair periods, customers are serviced at slower rates. The service time during this period are exponentially distributed with a rate of μ_R.
- If no arrivals occur during busy periods, the server waits in an inactive state, anticipating possible new arrivals for a random amount of time. The waiting server time follows an exponential distribution with a rate ofω.
- The model implements two types of vacations, mirroring the IEEE 802.16e Power Saving Classes:
    1. Type-I vacation: After waiting for new arrivals, the system enters an initial type-I vacation, which is exponentially distributed with a parameter of γ_1.
    2. Type-II vacation: After returning from this vacation, if the server finds no customers waiting in the queue, they enter a second, type-II vacation that follows exponential distribution with a parameter of γ_2. Otherwise, he returns to the regular state. If there are no customers left in the waiting queue after the type-II vacation, the server returns to the busy period to await new customers.
- During busy periods, the system may break down. In such cases, all present customers are removed from the system, and the system is sent for repair. The inter-arrival times between successive breakdowns are exponentially distributed with a rate of ϱ. Repair times follow an exponential distribution with a rate of ψ.

- When customers arrive during regular busy periods, type-I, type-II vacations, or repair periods, they decide whether to enter the system or balk. The balking probability depends on the state of the system. During regular busy periods, the balking probability is $1-\theta_B$. During type-I and type-II vacations, the balking probability is $1-\theta_V$. During repair periods, the balking probability is $1-\theta_R$.

- During type-I, type-II vacation, and repair periods, incoming customers may lose their patience and quit the system if they are not served. The impatience times follow an exponential distribution with a parameter of ξ.

All the variables associated with customers and server statuses are mutually independent. Figure 4.1 shows the state transition diagram of our system. Table 4.1 summarizes the parameters and variables of our model.

Variable	Description	Value or distribution
λ	Arrival rate of customers	Poisson process
μ_B	Service rate during regular busy periods	Exponential distribution
μ_R	Service rate during repair periods	Exponential distribution
ω	Waiting server rate	Exponential distribution
γ_1	Type-I vacation rate	Exponential distribution
γ_2	Type-II vacation rate	Exponential distribution
ϱ	Catastrophe rate	Exponential distribution
ψ	Repair rate	Exponential distribution
θ_B	Joining probability during regular busy periods	Constant value
θ_V	Joining probability during type-I and type-II vacations	Constant value
θ_R	Joining probability during repair periods	Constant value
ξ	Impatience rate	Exponential distribution

Table 4.1: Parameters and variables of the $M/M/1/DV$ queueing system with catastrophes

4.5 Model analysis

The behavior of the system is modeled as a two-dimensional continuous-time Markov process $\{(J(t), L(t)); t \geq 0\}$, where $J(t)$ denotes the server status and $L(t)$ represents the number of customers in the system at time t. The state space is defined as $\Omega = \{(j,n): j \in \{0,1,2,3\}, n \geq 0\}$, where the server state $J(t)$ is characterized as follows:

$$J(t) = \begin{cases} 0, & \text{the system is in type-II vacation period;} \\ 1, & \text{the system is in normal busy period;} \\ 2, & \text{the system is in type-I vacation period;} \\ 3, & \text{the system is in breakdown/repair period.} \end{cases}$$

Let $\pi_{j,n} = \lim_{t\to\infty} P(J(t) = j, L(t) = n)$ denote the steady-state probabilities for each state $(j,n) \in \Omega$. The state transition diagram illustrated in Figure 4.1.

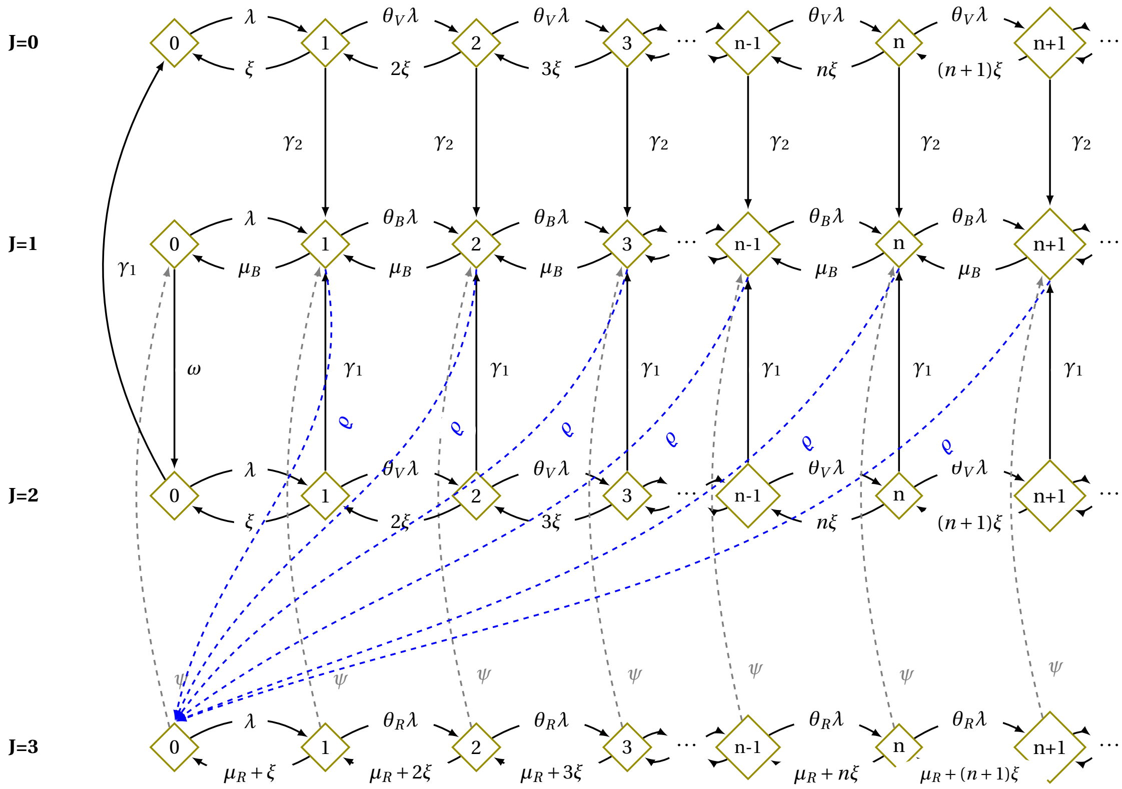

Figure 4.1: State transition diagram of the M/M/1/DV queueing model with catastrophes

4.5.1 Matrix Geometric solution

Consider the infinitesimal generator **Q**, which is defined as:

$$\mathbf{Q} = \begin{pmatrix} \Gamma_0 & \Lambda_0 & 0 & & & \\ \Psi_1+\Omega & \Gamma_1 & \Lambda_1 & & & \\ \Omega & \Psi_2 & \Gamma_2 & \Lambda_1 & & \\ \vdots & & \ddots & \ddots & \ddots & \\ \Omega & & & \Psi_N & \Gamma_N & \Lambda_1 \\ \vdots & & & & \ddots & \ddots \end{pmatrix}.$$

In this structure, the Quasi-Birth-Death (QBD) process is characterized by the matrices Γ_n and Ψ_n. These matrices are approximations for a large number of customers N where $n \geq N$. Each sub-matrix of the matrix Q is defined as follows:

$$\Gamma_0 = \begin{pmatrix} -\lambda & & & \\ & -(\lambda+\omega) & \omega & \\ \gamma_1 & & -(\lambda+\gamma_1) & \\ & \psi & & -(\lambda+\psi) \end{pmatrix},$$

$$\Psi_1 = \begin{pmatrix} \xi & & & \\ & \mu_B & & \\ & & \xi & \\ & & & \mu_R+\xi \end{pmatrix},$$

$$\Omega = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \varrho \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \Psi_n = \begin{pmatrix} n\xi & & & \\ & \mu_B & & \\ & & n\xi & \\ & & & \mu_R + n\xi \end{pmatrix}, \quad 2 \le n \le N-1,$$

$$\Psi_N = \begin{pmatrix} N\xi & & & \\ & \mu_B & & \\ & & N\xi & \\ & & & \mu_R + N\xi \end{pmatrix}, \quad n \ge N$$

$$\Lambda_0 = \begin{pmatrix} \lambda & & & \\ & \lambda & & \\ & & \lambda & \\ & & & \lambda \end{pmatrix}, \Lambda_1 = \begin{pmatrix} \theta_V\lambda & & & \\ & \theta_B\lambda & & \\ & & \theta_V\lambda & \\ & & & \theta_R\lambda \end{pmatrix},$$

$$\Gamma_n = \begin{pmatrix} \alpha_n & \gamma_2 & & \\ & \beta_n & & \\ & \gamma_1 & \kappa_n & \\ & \psi & & \eta_n \end{pmatrix}, \quad 1 \le n \le N-1, \quad \Gamma_N = \begin{pmatrix} \alpha_N & \gamma_2 & & \\ & \beta_N & & \\ & \gamma_1 & \kappa_N & \\ & \psi & & \eta_N \end{pmatrix}, \quad n \ge N,$$

where $\alpha_n = -(\theta_V\lambda + n\xi + \gamma_2), \kappa_n = -(\theta_V\lambda + n\xi + \gamma_1), \beta_n = -(\theta_B\lambda + \mu_B + \varrho), \eta_n = -(\theta_R\lambda + n\xi + \mu_R + \psi)$.

Given that the matrices have the same structure for $n \ge N$, we can apply the matrix-geometric method to obtain the stationary probabilities. This method is used in the analysis of stochastic models, particularly queueing models. In these models, the state space can be partitioned into levels such that transitions can occur from one level to either the same level or adjacent levels. This method provides a systematic way to compute the stationary distribution of the underlying Markov chain.

4.5.2 Stability condition

Theorem 4.1. *[19] The necessary and sufficient condition for the process to be positive recurrent is given by the following inequality:*

$$\frac{\lambda(\theta_B\psi + \theta_R\varrho)}{\psi\mu_B + \varrho(\mu_R + N\xi)} < 1. \tag{4.1}$$

Proof. The stability condition, which is both a sufficient and necessary condition for such a QBD, can be derived from the ergodic condition, as stated in [76]:

$$z\Lambda_1 e_n < z\Psi_N e_n,$$

where the row vector $z = [z_1, z_2, z_3, z_4]$ is the invariant probability vector. This vector is obtained as the unique solution to the equation $z(\Omega + \Lambda_N + \Gamma_N + \Psi_N) = 0$ such that $ze_n = 1$. Here, e_n is a column vector of ones and 0 is an n row vector. By solving the above equations, we find

$$\begin{cases} z_1 = 0, \\ z_2 = \frac{\psi}{\psi+\varrho}, \\ z_3 = 0, \\ z_4 = \frac{\varrho}{\psi+\varrho}. \end{cases}$$

Then, we have $z = [z_1, z_2, z_3, z_4] = [0, \frac{\psi}{\psi+\varrho}, 0, \frac{\varrho}{\psi+\varrho}]$. By substituting this result into the ergodic condition formula, we obtain the result in Equation (4.1). $\square$

4.5.3 Stationary probability distribution

Consider $\Pi = [\Pi_0, \Pi_1, \Pi_2, \dots]$ as the steady-state probability vector for the generator matrix $\mathbf{Q}$. Each Π_n is defined as $\Pi_n = [\pi_{0,n}, \pi_{1,n}, \pi_{2,n}, \pi_{3,n}]$. The following recursive relation is applied:

$$\Pi_n = \Pi_N R^{n-N}, \quad n \geq N. \tag{4.2}$$

Here, Π_n is referred to as the geometric probability vector. The rate matrix R is obtained as the minimal non-negative solution to the matrix quadratic equation:

$$\Lambda_1 + R\Gamma_N + R^2\Psi_N = 0. \tag{4.3}$$

The Quasi-Birth-Death (QBD) process is positive recurrent if the eigenvalues of R are within the unit disk, i.e., the spectral radius $Sp(R) < 1$. However, deriving an explicit expression for the rate matrix, denoted as R, from the nonlinear Equation (4.3) poses a significant challenge.

In accordance with the studies conducted by [76] and [65], an approximate solution for R can be obtained by considering the sequence $\{R_k\}_{k\geq 0}$, which is defined as follows:

$$R_0 = 0, \quad \text{and} \quad R_{k+1} = (\Lambda_1 + R_k^2\Psi_N)(-\Gamma_N)^{-1}, \quad k \geq 0. \tag{4.4}$$

The iterative procedure continues until the convergence criterion $\|R_{k+1} - R_k\| < \varepsilon$ is satisfied, where ε represents the tolerance threshold.

The stationary probability vector Π satisfies the global balance equation $\Pi\mathbf{Q} = 0$, which leads to the system of equations:

$$\begin{aligned}
\Pi_0\Gamma_0 + \Pi_1(\Psi_1 + \Omega) + \sum_{n=2}^{\infty} \Pi_n\Omega &= 0, \quad n = 0 && (4.5)\\
\Pi_0\Lambda_0 + \Pi_1\Gamma_1 + \Pi_2\Psi_2 &= 0, \quad n = 1 && (4.6)\\
\Pi_{n-1}\Lambda_1 + \Pi_n\Gamma_n + \Pi_{n+1}\Psi_{n+1} &= 0, \quad 2 \leq n \leq N-1 && (4.7)\\
\Pi_{n-1}\Lambda_1 + \Pi_n\Gamma_N + \Pi_{n+1}\Psi_N &= 0, \quad n \geq N. && (4.8)
\end{aligned}$$

The normalizing condition is given by:

$$\sum_{n=0}^{\infty} \Pi_n e_n = 1, \tag{4.9}$$

where e_n represents the n-th element of a column vector e of appropriate dimension, with all elements equal to one. That is, $e_n = 1$ for all n.

For $n = N+1$, equation (4.8) becomes $\Pi_N\Lambda_1 + \Pi_{N+1}\Gamma_N + \Pi_{N+2}\Psi_N = 0$. Substituting (4.2) into this equation, we obtain the quadratic matrix equation:

$$\Lambda_1 + R\Gamma_N + R^2\Psi_N = 0$$

which defines the rate matrix R (4.3) and consequently (4.4).
Then, from equation (4.8) at $n = N$, we have $\Pi_{N-1}\Lambda_1 + \Pi_N\Gamma_N + \Pi_{N+1}\Psi_N = 0$. Substituting $\Pi_{N+1} = \Pi_N R$ and isolating Π_{N-1} yields:

$$\Pi_{N-1} = -\Pi_N(\Gamma_N + R\Psi_N)\Lambda_1^{-1} = \Pi_N\mathscr{H}_1,$$

where $\mathscr{H}_1 = -(\Gamma_N + R\Psi_N)\Lambda_1^{-1}$. By repeatedly applying this backward substitution into equation (4.9) for $n = N-1, N-2, \dots, 2$, we find the general relation $\Pi_n = \Pi_N\mathscr{H}_{N-n}$, where the matrices $\mathscr{H}_n$ are defined as:

$$\mathscr{H}_n = -(\mathscr{H}_{n-1}\Gamma_{N-n+1} + \mathscr{H}_{n-2}\Psi_{N-n+2})\Lambda_1^{-1}, \quad n = 2, 3, \dots, N-1.$$

Substituting $\Pi_1 = \Pi_N\mathscr{H}_{N-1}$ and $\Pi_2 = \Pi_N\mathscr{H}_{N-2}$ into equation (4.6), we obtain:

$$\Pi_0 = -\Pi_N(\mathscr{H}_{N-1}\Gamma_1 + \mathscr{H}_{N-2}\Psi_2)\Lambda_0^{-1} = \Pi_N\mathscr{H}_N.$$

This confirms $\mathscr{H}_N = -(\mathscr{H}_{N-1}\Gamma_1 + \mathscr{H}_{N-2}\Psi_2)\Lambda_0^{-1}$.
Substituting $\Pi_n = \Pi_N\mathscr{H}_{N-n}$ and the geometric tail (4.2) into equation (4.5) leads to:

$$\Pi_N\left[\mathscr{H}_N\Gamma_0 + \mathscr{H}_{N-1}\Psi_1 + \sum_{n=1}^{N-1}\mathscr{H}_{N-n}\Omega + (I-R)^{-1}\Omega\right] = 0 \tag{4.10}$$

Finally, the normalization condition (4.9) gives:

$$\Pi_N\left[\sum_{n=1}^{N}\mathscr{H}_n + (I-R)^{-1}\right]e_n = 1. \tag{4.11}$$

Solving the system formed by equations (4.10) and (4.11) provides the value of Π_N. Finally, using (4.2) and $\Pi_n = \Pi_N\mathscr{H}_{N-n}$, for $n = \overline{0, N-1}$, we get the remaining Π_n which completes the determination of all stationary probabilities.

4.6 Performance measures

- Probability of the system in busy period:

$$\mathscr{P}_B = \sum_{n=0}^{\infty}\pi_{1,n}.$$

- Probability of the system in vacations: Type-II vacation:

$$\mathscr{P}_{V0} = \sum_{n=0}^{\infty}\pi_{0,n},$$

  and Type-I vacation:

$$\mathscr{P}_{V2} = \sum_{n=0}^{\infty}\pi_{2,n}.$$

- The probability that the servers are idle during busy period:

$$\mathscr{P}_I = \pi_{1,0}.$$

- Probability of the system in vacation period :

$$\mathscr{P}_V = P_{V0} + P_{V2}.$$

- The probability that the servers are in repair process:

$$\mathscr{P}_R = \sum_{n=0}^{\infty} \pi_{3,n}.$$

- Probability of system reliability :

$$\mathscr{P}_{re} = 1 - P_R.$$

- Mean system length:

$$\mathscr{L}_s = \sum_{n=0}^{\infty} n(\pi_{0,n} + \pi_{1,n} + \pi_{2,n} + \pi_{3,n}).$$

- The mean number of customers served per unit time:

$$\mathscr{E} = \mu_B \sum_{n=0}^{\infty} n\pi_{1,n} + \mu_R \sum_{n=0}^{\infty} n\pi_{3,n}.$$

- Expected delay of customers:

$$\mathscr{E}_{delay} = \mathscr{L}_s / \mathscr{E}.$$

- The average balking rate :

$$\mathscr{B}_r = \lambda \sum_{n=0}^{\infty} (1-\theta_V)(\pi_{0,n} + \pi_{2,n}) + \lambda \sum_{n=0}^{\infty} (1-\theta_B)\pi_{1,n} + \lambda \sum_{n=0}^{\infty} (1-\theta_R)\pi_{3,n}.$$

- The average reneging rate :

$$\mathscr{R}_r = \xi \left(\sum_{n=0}^{\infty} n\pi_{0,n} + \sum_{n=0}^{\infty} n\pi_{2,n} + \sum_{n=0}^{\infty} n\pi_{3,n} \right).$$

4.7 Numerical analysis

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

4.7.1 Cost analysis

To develop a cost model, we define the governing unit cost elements as:

- C_1 : Cost per unit time per customer served in the system: This represents the expenses incurred for for providing service to each customer in the system at any given time.

- C_2 : Cost per unit time for delayed service : This signifies the cost associated as a penalty cost for giving delayed service per unit time per customer.

- C_3 : Cost per unit time when a customer balks: It represents the cost attributed to customers who are refraining from joing the queue due to dissatisfaction.

- C_4 : 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_5 : 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_6 : Cost per service per unit time when the servers are on vacation: This cost pertains to the cost during vacation periods when the servers are not actively engaged in customer service.

Here, it's noteworthy that the cost structure and notation are in accordance with well-known results. Further, the arbitrary values of the parameters are chosen so that the steady state condition is verified. 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

$$Cost = C_1\mathscr{E} + C_2\mathscr{E}_{delay} + C_3\mathscr{B}_R + C_4\mathscr{R}_r + C_5\mathscr{P}_I + C_6\mathscr{P}_V.$$

4.7.2 Numerical illustrations

To study the qualitative aspects of the queueing model under consideration, numerical computations have been carried out. For computational purpose, the values are taken as $N = 50, \epsilon = 10^-5$. The other parameters are $\lambda = 2.5, \omega = 0.15, \gamma_1 = 0.4, \gamma_2 = 0.2$, $\psi = 0.43, \xi = 0.15, \varrho = 0.13$, $\theta_V = 0.1, \theta_B = 0.82, \theta_R = 0.14, \mu_B = 4.5$, and $\mu_R = 1.2$ unless otherwise they have a particular mention in the graphs and tables. The cost parameters are assumed as $C_1 = 50$, $C_2 = 150$, $C_3 = 35$, $C_4 = 145$, $C_5 = 100$ and $C_6 = 50$.

Table 4.2: Effect of μ_B on different performance measures

μ_B	$\mathscr{P}_I$	$\mathscr{P}_{re}$	$\mathscr{B}_r$	$\mathscr{R}_r$	$\mathscr{L}_s$	$\mathscr{E}_{delay}$	$Cost$
4.0	0.313468	0.878196	0.736787	0.0429842	1.07951	0.329072	284.843
4.5	0.347131	0.890046	0.725204	0.04482	0.953687	0.313188	274.788
5.0	0.375165	0.899913	0.715559	0.0463487	0.866015	0.301284	267.88
5.5	0.398758	0.908218	0.707441	0.0476353	0.801863	0.292064	262.922
6.0	0.418835	0.915285	0.700533	0.0487302	0.753089	0.284732	259.235
6.5	0.436101	0.921363	0.694592	0.0496717	0.71486	0.278773	256.412
7.0	0.451093	0.92664	0.689434	0.0504893	0.684149	0.273843	254.198
7.5	0.464223	0.931262	0.684917	0.0512053	0.65897	0.2697	252.425
8.0	0.475813	0.935341	0.680929	0.0518373	0.637974	0.266175	250.98

Table 4.3: Effect of μ_R on different performance measures

μ_R	$\mathscr{P}_I$	$\mathscr{P}_{re}$	$\mathscr{B}_r$	$\mathscr{R}_r$	$\mathscr{L}_s$	$\mathscr{E}_{delay}$	$Cost$
1.0	0.418586	0.915197	0.700619	0.0493744	0.758068	0.287052	259.446
1.5	0.41915	0.915396	0.700425	0.0479192	0.74686	0.281784	258.989
2.0	0.419562	0.915541	0.700283	0.0468625	0.738807	0.277903	258.703
2.5	0.419877	0.915652	0.700175	0.0460551	0.732697	0.27491	258.509
3.0	0.420127	0.91574	0.700089	0.0454154	0.72788	0.272522	258.37
3.5	0.420331	0.915812	0.700018	0.0448948	0.723973	0.27057	258.266
4.0	0.420501	0.915872	0.69996	0.0444619	0.720734	0.268942	258.186

Table 4.4: Effect of λ on different performance measures

λ	$\mathscr{P}_I$	$\mathscr{P}_{re}$	$\mathscr{B}_r$	$\mathscr{R}_r$	$\mathscr{L}_s$	$\mathscr{E}_{delay}$	$Cost$
1	0.471062	0.955828	0.326684	0.0447039	0.478998	0.573222	209.72
1.5	0.442726	0.933206	0.449937	0.0448806	0.605169	0.426286	214.93
2	0.398004	0.911208	0.581926	0.0448465	0.759318	0.35477	237.877
2.5	0.347131	0.890046	0.725204	0.04482	0.953687	0.313188	274.788
2.7	0.326128	0.881853	0.785798	0.0448201	1.04628	0.301153	293.746
3	0.294507	0.869893	0.880133	0.0448348	1.20544	0.286366	327.434

Table 4.5: Effect of impatient rate ξ on different performance measures

ξ	$\mathscr{P}_I$	$\mathscr{P}_{re}$	$\mathscr{B}_r$	$\mathscr{R}_r$	$\mathscr{L}_s$	$\mathscr{E}_{delay}$	$Cost$
0.1	0.347699	0.889374	0.719735	0.0311534	0.973096	0.316128	274.594
0.15	0.347131	0.890046	0.725204	0.04482	0.953687	0.313188	274.788
0.2	0.346328	0.890687	0.731074	0.0578242	0.938326	0.311039	275.237
0.25	0.345358	0.891307	0.73723	0.070376	0.92565	0.309435	275.851
0.3	0.344266	0.891911	0.743587	0.0825992	0.914839	0.308217	276.576
0.35	0.343086	0.892502	0.750086	0.094571	0.905374	0.30728	277.379

Table 4.6: Effect of repair rate ψ on different performance measures

ψ	$\mathscr{P}_I$	$\mathscr{P}_{re}$	$\mathscr{B}_r$	$\mathscr{R}_r$	$\mathscr{L}_s$	$\mathscr{E}_{delay}$	$Cost$
0.2	0.307555	0.789771	0.885607	0.0546818	0.947565	0.334796	269.353
0.4	0.344221	0.88274	0.736885	0.0455165	0.953196	0.314613	274.378
0.6	0.358655	0.918763	0.679315	0.0421487	0.955744	0.30786	276.414
0.8	0.366449	0.937885	0.64879	0.0404585	0.95728	0.304602	277.505
1.0	0.371363	0.949736	0.629893	0.0394691	0.958334	0.302739	278.18
1.2	0.374764	0.957799	0.617052	0.0388336	0.959113	0.301563	278.637

Table 4.7: Effect of catastrophe rate ϱ on different performance measures

ϱ	$\mathscr{P}_I$	$\mathscr{P}_{re}$	$\mathscr{B}_r$	$\mathscr{R}_r$	$\mathscr{L}_s$	$\mathscr{E}_{delay}$	$Cost$
0.12	0.347131	0.890046	0.725204	0.04482	0.953687	0.313188	274.788
0.2	0.332013	0.842358	0.802429	0.0487205	0.925148	0.325504	267.857
0.22	0.328026	0.829883	0.822622	0.0497378	0.917959	0.328963	266.142
0.3	0.313341	0.784294	0.896389	0.0534449	0.89266	0.34253	260.229
0.32	0.309953	0.773858	0.913269	0.0542911	0.887084	0.345854	258.957
0.4	0.297378	0.735379	0.975485	0.0574031	0.867223	0.358875	254.536

Table 4.8: The impact of service rate μ_B and vacation rates γ_1, γ_2

μ_B	γ_1	$\mathscr{L}_s$	$\mathscr{E}_{delay}$	γ_2	$\mathscr{L}_s$	$\mathscr{E}_{delay}$
	γ_1=1.0	1.59072	0.376243	γ_2=0.1	1.60682	0.39994
$\mu_B = 3.0$	γ_1=2.0	1.59476	0.381433	γ_2=0.5	1.58448	0.363433
	γ_1=3.0	1.59687	0.386271	γ_2=1.0	1.58427	0.359773
	γ_1=1.0	0.843026	0.283603	γ_2=0.1	0.93051	0.340841
$\mu_B = 5.0$	γ_1=2.0	0.863864	0.296151	γ_2=0.5	0.791634	0.252619
	γ_1=3.0	0.881069	0.30781	γ_2=1.0	0.777306	0.243752
	γ_1=1.0	0.648936	0.249446	γ_2=0.1	0.771871	0.32861
$\mu_B = 7.0$	γ_1=2.0	0.678447	0.266808	γ_2=0.5	0.703246	0.282917
	γ_1=3.0	0.703246	0.282917	γ_2=1.0	0.550714	0.194275

Table 4.9: Effect of waiting server rate ω on different performance measures

ω	$\mathscr{P}_I$	$\mathscr{P}_{re}$	$\mathscr{B}_r$	$\mathscr{R}_r$	$\mathscr{L}_s$	$\mathscr{E}_{delay}$	$Cost$
0.20	0.323873	0.894798	0.738667	0.0522339	0.977775	0.334073	273.415
0.25	0.303535	0.898953	0.750438	0.0587166	0.998838	0.353768	272.429
0.30	0.285601	0.902617	0.760818	0.0644333	1.01741	0.372373	271.745
0.35	0.269667	0.905873	0.770041	0.0695122	1.03391	0.389976	271.298
0.40	0.255418	0.908784	0.778288	0.0740543	1.04867	0.406655	271.039
0.45	0.242599	0.911403	0.785708	0.0781405	1.06195	0.422482	270.93

Table 4.10: Effect of joining rates $\theta_B, \theta_V, \theta_R$ on different performance measures

	$\mathscr{P}_I$	$\mathscr{P}_{re}$	$\mathscr{B}_r$	$\mathscr{R}_r$	$\mathscr{L}_s$	$\mathscr{E}_{delay}$	$Cost$
θ_B =0.2	0.474014	0.935973	2.02771	0.0667041	0.568342	0.359216	273.543
θ_B =0.5	0.462825	0.932004	1.59723	0.0659259	0.600525	0.329938	264.455
θ_B =0.9	0.450221	0.927534	1.16932	0.0650494	0.641182	0.302906	258.536
θ_V =0.1	0.434407	0.921925	0.702191	0.0639496	0.699418	0.275802	256.921
θ_V =0.3	0.431651	0.919222	0.66339	0.0944768	0.887889	0.320396	277.723
θ_V =0.5	0.427037	0.916812	0.635001	0.126182	1.09486	0.354531	301.545
θ_R =0.1	0.434595	0.921992	0.70993	0.0632388	0.695205	0.275083	256.57
θ_R =0.3	0.433522	0.921612	0.671126	0.0673137	0.71976	0.279014	258.877
θ_R =0.5	0.432104	0.921109	0.631944	0.0727946	0.754199	0.283609	262.792

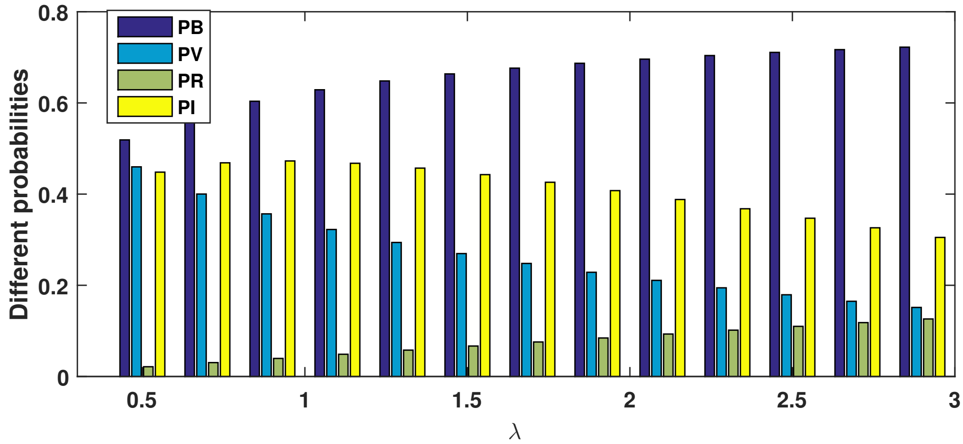

Figure 4.2: The effect of λ on different probabilities

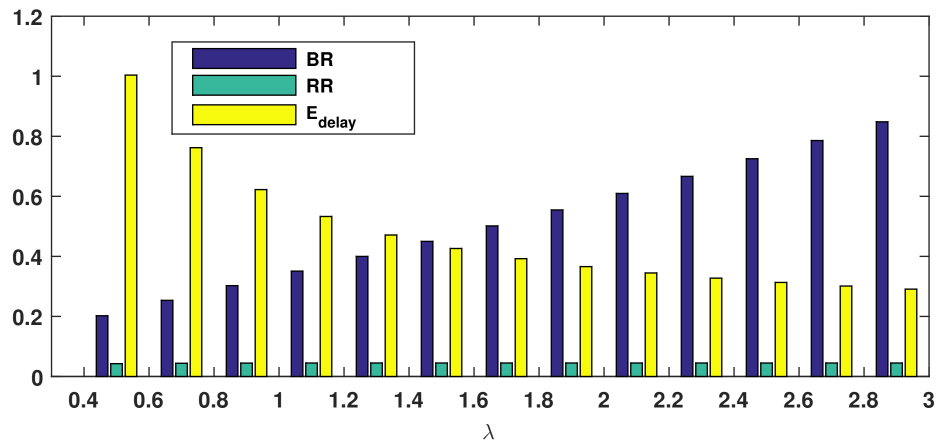

Figure 4.3: The effect of λ on expected delay, expected balking and reneging rates

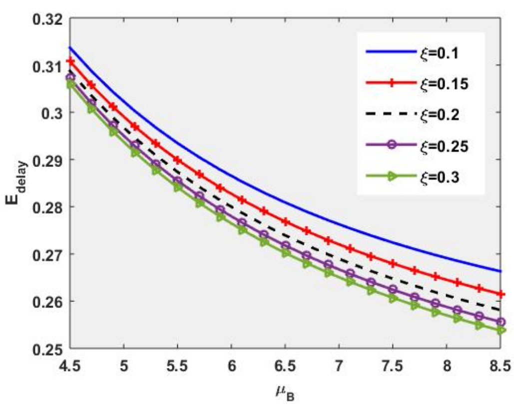

Figure 4.4: The effect of μ_B on expected delay for different reneging rates ξ

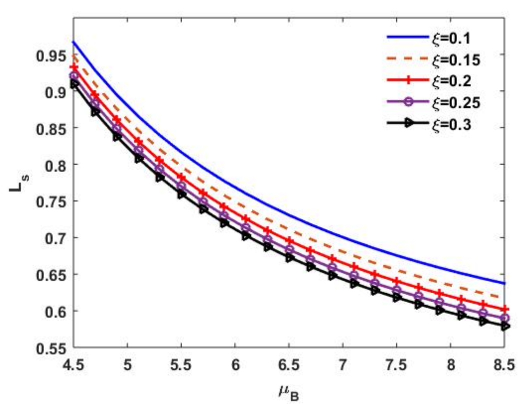

Figure 4.5: The effect of μ_B on expected system length $\mathscr{L}_s$

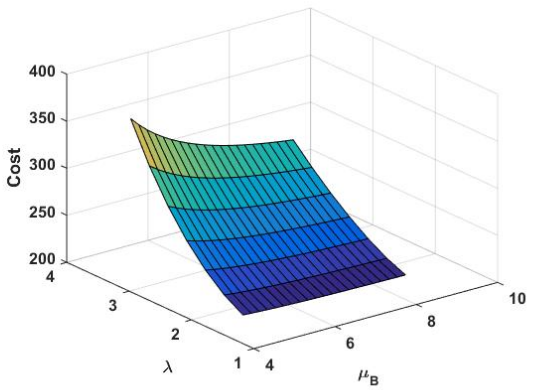

Figure 4.6: The effect of λ and μ_B on the expected cost

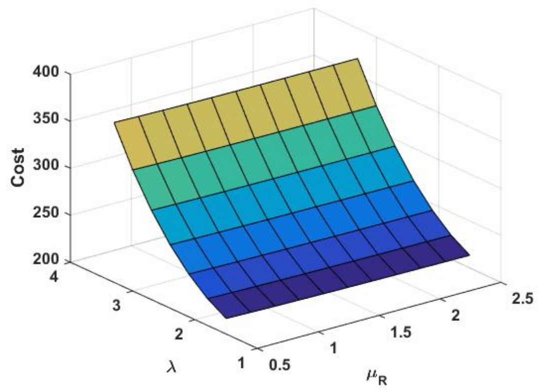

Figure 4.7: The effect of λ and μ_R on the expected cost

4.7.3 Discussion of results

The analysis of our model for the IEEE 802.16e power saving mechanism, modeled as a vacation system with waiting servers, impatience, and catastrophes, reveals several key insights into system behavior and performance. We discuss the results presented in Tables 4.2-4.10 along with Figures 4.2-4.7.

4.7.4 Combined effect of service rate (μ_B) and vacation rates (γ_1, γ_2)

Tables 4.2 and 4.8, along with Figures 4.4 and 4.5, provide crucial insights into how the service rate (μ_B) and vacation rates (γ_1, γ_2) jointly affect the performance of the IEEE 802.16e Power Saving Mechanism, modeled as a vacation system with waiting servers, impatience, and catastrophes. Our analysis reveals several significant trends and interactions:

1. **Impact of service rate (μ_B):**

    - Higher μ_B results in increased idleness probability ($\mathscr{P}_I$) and system reliability ($\mathscr{P}_{re}$), enabling more frequent mobile station sleep modes and enhanced overall system stability.
    - The rise in μ_B consistently reduces mean system size ($\mathscr{L}_s$) and expected delay ($\mathscr{E}_{delay}$), indicating fewer queued packets and shorter transmission times, as validated by Figures 4.4 and 4.5.
    - The average balking rate ($\mathscr{B}_r$) declines with increasing μ_B, signaling fewer dropped packets. Simultaneously, a slight increase in the average reneging rate ($\mathscr{R}_r$) suggests more system entries with minimal delay impact.
    - Total expected cost significantly decreases with higher μ_B, highlighting improved resource utilization and potential energy savings.

    These observations suggest that higher service rates lead to more efficient power-saving mechanisms in IEEE 802.16e. Faster service rates allow for quicker data processing during active periods, reducing queue length and packet delays, potentially allowing for longer or more frequent sleep periods without compromising performance.

2. Impact of initial vacation rate (γ_1):

    For all service rates μ_B, increasing γ_1 leads to increases in both $\mathscr{L}_s$ and $\mathscr{E}_{delay}$. This trend is more pronounced at higher service rates. For example, at $\mu_B = 7.0$, the increase in $\mathscr{L}_s$ is more significant compared to $\mu_B = 3.0$.

    These findings suggest that longer initial sleep periods (higher γ_1) lead to more accumulated data in the system. In the context of IEEE 802.16e, extended initial sleep windows may result in increased buffer occupancy and longer packet delays, necessitating a balance between power-saving benefits and performance impacts.

3. Impact of subsequent vacation rate (γ_2):

   The effect of γ_2 varies with the service rate μ_B. At lower service rates (e.g., $\mu_B = 3.0$), increasing γ_2 slightly decreases $\mathscr{L}_s$ and $\mathscr{E}_{delay}$. However, at higher service rates (e.g., $\mu_B = 7.0$), the impact of increasing γ_2 is more pronounced, with larger decreases in both $\mathscr{L}_s$ and $\mathscr{E}_{delay}$.

   These observations highlight the importance of subsequent sleep period duration in balancing power saving with queue management. For IEEE 802.16e, this suggests that the effectiveness of adjusting subsequent sleep windows depends on the system's service capacity.

4. Interaction between μ_B, γ_1, and γ_2:

   The impact of changing γ_1 or γ_2 becomes more significant as μ_B increases. Further, at higher μ_B, the system shows greater sensitivity to changes in vacation rates.

   This interaction demonstrates the complex relationship between service capacity and sleep scheduling in the IEEE 802.16e power-saving mechanism. It suggests that systems with higher service rates have more flexibility in adjusting their sleep patterns without severely impacting performance.

5. Power saving class implications:

   - Type-I class: Increasing γ_1 at higher μ_B enables more effective exponential sleep window expansion.
   - Type-II class: Variable γ_2 impact suggests optimal fixed sleep window depends on service capacity.

6. Power saving mechanism insights:

   - Higher μ_B creates more power conservation opportunities for mobile stations.
   - Sleep scheduling effectiveness depends critically on dynamic service conditions.
   - Power-saving mechanisms must balance sleep periods against system performance metrics.
   - Distinct Type-I and Type-II class behaviors demand tailored optimization strategies.

4.7.5 Effect of service rate during repair process (μ_R)

Table 4.3 illustrates the impact of the service rate during the repair process or breakdown period (μ_R) on various performance measures of the IEEE 802.16e power-saving mechanism. As μ_R increases, we observe several subtle but notable effects:

- Marginal increases in system idleness ($\mathscr{P}_I$) and reliability ($\mathscr{P}_{re}$) suggest improved data processing during repair periods.
- Slight decrease in balking rate indicates reduced packet dropping.
- Substantial decrease in reneging rate shows improved service during system issues.
- Reduced mean system length ($\mathscr{L}_s$) and expected delay ($\mathscr{E}_{delay}$) demonstrate enhanced system efficiency.
- Total expected cost decreases from 259.446 to 258.186, indicating modest cost savings.
- Improvements during repair periods show limited impact compared to busy period service rate enhancements due to:
  - Less frequent and shorter breakdown periods.
  - Reduced system capacity during repairs.
  - Power-saving mechanism's design prioritizing normal operation efficiency.

4.7.6 Effect of (λ) and (μ_B, μ_R) on IEEE 802.16e Power-Saving Mechanism

This analysis explores the combined effects of arrival rate (λ), busy period service rate (μ_B), and repair period service rate (μ_R) on the IEEE 802.16e power-saving mechanism, as demonstrated in Table 4.4 and Figures 4.2, 4.3, 4.6, and 4.7:

1. As λ increases, we observe:

   - Idleness probability ($\mathscr{P}_I$) and vacation probability ($\mathscr{P}_V$) decrease (Figure 4.2), indicating fewer opportunities for the Mobile Station (MS) to enter sleep mode. In addition, server busyness ($\mathscr{P}_B$) rises (Figure 4.2), reflecting more frequent data transmissions. The rate of increase in $\mathscr{P}_B$ slows at higher λ values, suggesting a saturation point.
   - System reliability ($\mathscr{P}_{re}$) declines, while repair probability ($\mathscr{P}_R$) slightly increases (Figure 4.2), indicating more frequent system breakdowns under higher loads.
   - Mean system size ($\mathscr{L}_s$) increases substantially, indicating longer queues and higher congestion. While, expected delay ($\mathscr{E}_{delay}$) decreases (Figure 4.3), possibly due to higher overall packet processing rates.
   - Average balking rate ($\mathscr{B}_r$) increases significantly (Figure 4.3), indicating more packets being dropped or redirected. However, the average reneging rate ($\mathscr{R}_r$) remains relatively stable across different λ values (Figure 4.3).

These observations demonstrate a fundamental trade-off between system activity and power-saving opportunities, with non-linear relationships suggesting the need for dynamic power-saving strategies.

2. Figures 4.6 and 4.7 illustrate the combined effects of λ, μ_B, and μ_R on the total expected cost:

   - Arrival rate (λ) effect:
   $$\frac{\partial \text{Cost}}{\partial \lambda} > 0.$$
   Higher network traffic leads to increased system costs, potentially due to higher energy consumption and resource utilization.
   - Busy period service rate (μ_B) effect:
   $$\frac{\partial \text{Cost}}{\partial \mu_B} < 0.$$
   Improving the service rate during busy periods can effectively mitigate the cost increase caused by higher arrival rates.
   - Repair period service rate (μ_R) effect:
   $$\left|\frac{\partial \text{Cost}}{\partial \mu_R}\right| < \left|\frac{\partial \text{Cost}}{\partial \mu_B}\right|.$$
   While improving repair-time service rates is beneficial, it's less critical than enhancing busy-time efficiency for overall cost reduction.
   - Interactions:
   $$\frac{\partial^2 \text{Cost}}{\partial \lambda \partial \mu_B} < 0 \quad \text{and} \quad \frac{\partial^2 \text{Cost}}{\partial \lambda \partial \mu_R} < 0.$$
   Enhancing service efficiency becomes particularly crucial during high-traffic periods, with μ_B having a more pronounced effect than μ_R.

3. Based on these observations and mathematical interpretations, we can draw several implications:

   (a) Load balancing and traffic management strategies are crucial, given the system's pronounced sensitivity to arrival rates and their direct impact on power consumption dynamics.

   (b) Prioritizing improvements in the busy period service rate (μ_B) over the repair period service rate (μ_R) offers the most significant potential for enhancing overall system performance and energy efficiency.

   (c) The observed non-linear relationships necessitate the development of adaptive power-saving strategies capable of dynamically responding to fluctuating arrival rates and service conditions.

(d) While increased arrival rates correlate with reduced average delay, this optimization is counterbalanced by heightened packet loss due to balking, underscoring the imperative of designing power-saving mechanisms that maintain a delicate equilibrium between performance and service quality.

4.7.7 Effects of reneging parameter (ξ) and service rate (μ_B)

The analysis of Table 4.5 and Figures 4.4 and 4.5 reveals significant insights into the behavior of the IEEE 802.16e Power Saving Mechanism under varying conditions of customer impatience (ξ) and service rate (μ_B):

1. As the reneging parameter ξ increases, both average balking ($\mathscr{B}_r$) and reneging ($\mathscr{R}_r$) rates rise significantly, as shown in Table 4.5. This increased impatience leads to fewer data packets entering or remaining in the system, resulting in decreased mean system length ($\mathscr{L}_s$) and expected delay ($\mathscr{E}_{delay}$) as illustrated in Figures 4.4 and 4.5.

2. The system demonstrates a remarkable self-regulating mechanism. Increased ξ and μ_B both improve system efficiency, albeit through distinct approaches. While higher μ_B processes packets faster, increased ξ reduces system load by effectively dropping packets, serving as a natural control mechanism during high traffic periods.

3. The improved efficiency comes with notable consequences. Higher balking and reneging rates indicate potential data loss and reduced Quality of Service (QoS). This trade-off is reflected in the increasing system cost, driven by penalties associated with reneged customers as shown in Table 4.5.

4. The data in Table 4.5 reveals a marginal decrease in idle server probability ($\mathscr{P}_I$) and a slight improvement in system reliability ($\mathscr{P}_{re}$). This suggests base stations spend less time in deep sleep modes, maintaining a more active state to serve subscribers more responsively.

5. The IEEE 802.16e Power Saving Mechanism faces a critical challenge in balancing efficiency and reliability. While increasing μ_B enhances service quality by faster packet processing, higher ξ improves efficiency through packet discarding. System designers must carefully navigate these trade-offs, considering power-saving benefits against potential reliability issues.

6. The analysis highlights the potential for context-aware power management. By dynamically adjusting strategies based on service rates and customer patience levels, it becomes possible to optimize both power consumption and service quality, as evidenced by the intricate relationships in Table 4.5 and Figures 4.4 and 4.5.

4.7.8 Effect of repair rate (ψ)

Table 4.6 provides crucial insights into how the repair rate (ψ), varying from 0.2 to 1.2, affects the performance of the IEEE 802.16e Power Saving Mechanism, modeled as a vacation system with waiting servers, impatience, and catastrophes. Our analysis reveals several significant trends:

- As ψ increases, the probability of idle servers ($\mathscr{P}_I$) rises, accompanied by a significant improvement in system reliability ($\mathscr{P}_{re}$). This indicates the system's enhanced ability to recover quickly from network disruptions, spending more time in operational states and less in repair modes.

- The analysis reveals substantial improvements in critical performance metrics. Both balking ($\mathscr{B}_r$) and reneging ($\mathscr{R}_r$) rates decrease significantly, reflecting increased network stability. This suggests more consistent service delivery, and reduced data packet loss.

- Despite the improved stability, the system experiences a slight increase in mean system length ($\mathscr{L}_s$), indicating an enhanced capacity to handle data packets simultaneously. Notably, the expected delay ($\mathscr{E}_{delay}$) reduces, suggesting more efficient power-saving mechanism operations and smoother transitions between active and sleep states.

- The total expected system cost increases, representing the trade-off between improved performance and operational expenses. This reflects the resource investment required to achieve higher network reliability and reduced data loss.

- The improved repair rates offer nuanced benefits to different power saving classes:
  - Type-I vacation mechanisms can use more stable exponential sleep windows.
  - Type-II vacation strategies benefit from more predictable traffic patterns and more efficient fixed sleep window utilization.

- The findings underscore the critical role of repair rates in maintaining network resilience, highlighting the delicate balance between performance optimization and operational costs in mobile communication systems.

4.7.9 Effect of catastrophe rate (ϱ)

Table 4.7 provides critical insights into how the catastrophe rate (ϱ), varying from 0.12 to 0.4, affects the performance of the IEEE 802.16e Power Saving Mechanism, modeled as a vacation system with waiting servers, impatience, and catastrophes. Our analysis reveals several significant trends:

1. The probability of idle servers ($\mathscr{P}_I$) decreases, accompanied by a substantial drop in system reliability ($\mathscr{P}_{re}$). These trends indicate that higher ϱ values force the

system to spend more time in recovery states, effectively reducing its operational efficiency and availability.

2. Concurrent with system state changes, there are notable shifts in traffic metrics. The average balking rate ($\mathscr{B}_r$) and reneging rate ($\mathscr{R}_r$) both increase substantially, signaling decreased network stability. These increases reflect a network becoming less reliable, with more data packets either refusing to enter or prematurely leaving the system.

3. The system's performance metrics paint a concerning picture. Mean system length ($\mathscr{L}_s$) decreases, while expected delay ($\mathscr{E}_{delay}$) increases, suggesting a compromised power-saving mechanism. This indicates that the network becomes less efficient at managing data packets, with longer queuing times and reduced throughput.

4. Interestingly, the total expected cost decreases, likely due to reduced operational time during frequent catastrophic events. However, this cost reduction comes at a significant price of network reliability and performance degradation.

5. The catastrophe rate's effects extend to different power saving classes. Type-I vacation mechanisms face disrupted exponential sleep windows, while Type-II vacation strategies struggle to maintain consistent fixed sleep windows. The result is increased energy consumption and reduced power-saving efficiency.

6. The findings underscore the critical sensitivity of the IEEE 802.16e power-saving mechanism to catastrophic events. Higher catastrophe rates fundamentally compromise the network's ability to maintain stable operations, presenting significant challenges for reliable mobile communication systems.

4.7.10 Effect of waiting server rate (ω)

The waiting server rate ω significantly influences the performance of the IEEE 802.16e Power Saving Mechanism. From Table 4.9, we observe several key trends:

- As ω increases, $\mathscr{P}_I$ decreases. This occurs because a higher waiting server rate means the system spends less time in the idle state before transitioning to a vacation state. Consequently, the system becomes more responsive to incoming traffic, reducing overall idle time.

- Both $\mathscr{P}_{re}$ and $\mathscr{R}_r$ show slight increases with ω. This behavior can be attributed to the increased system activity. As the system spends less time idle, more customers accumulate in the queue during busy periods. This longer queue increases the likelihood that some customers will become impatient and leave before being served.

- The balking rate increases with ω. This is because a more active system (less idle time) is more likely to be found in a busy state when new customers arrive. Given that the balking probability is higher during busy periods than during vacations, this leads to an overall increase in balking.

- Both these metrics grow with ω. The primary reason is that a more active system (higher ω) allows more customers to enter before transitioning to a vacation state. This leads to larger queue sizes and, consequently, longer delays. Additionally, the increased reneging rate doesn't fully compensate for the higher influx of customers, resulting in net growth of the system size.

- Interestingly, the overall system cost decreases as ω increases. This suggests that the benefits of reduced idle time (improved energy efficiency) and increased throughput outweigh the negative effects of increased balking, reneging, and delays. The power-saving mechanism becomes more efficient at higher ω values, as it more quickly transitions to power-saving modes when no traffic is present.

4.7.11 Effect of joining probabilities ($\theta_B, \theta_V, \theta_R$)

Table 4.10 illustrates the impact of different joining rates on system performance. The reasons for the observed behaviors are as follows:

1. As joining probability increases during busy periods, (θ_B), the system experiences notable performance shifts. The idle probability ($\mathscr{P}_I$) and system reliability ($\mathscr{P}_{re}$) decrease slightly, indicating more efficient resource utilization. The system benefits from significantly reduced average balking rates ($\mathscr{B}_r$), suggesting more customers actively engage with the service. Interestingly, while mean system length ($\mathscr{L}_s$) increases with more customers, the expected delay ($\mathscr{E}_{delay}$) paradoxically decreases, potentially due to more streamlined service processes. The total expected cost reduction further validates the benefits of higher joining rates during busy periods.

2. The consequences of increased joining during vacation periods, (θ_V), are markedly different. Idle probability ($\mathscr{P}_I$) decreases as the system becomes more likely to have waiting customers. However, this comes with significant drawbacks: average reneging rates ($\mathscr{R}_r$) significantly increase, reflecting customer impatience during prolonged wait times. Mean system length and expected delay, $\mathscr{L}_s$ and $\mathscr{E}_{delay}$, increase substantially, indicating increased congestion. Additional, the total expected cost rises dramatically, suggesting that higher vacation period joining probabilities can be detrimental to system performance and potentially compromise the energy-saving mechanisms.

3. Joining during repair periods (θ_R) shows more subtle effects. Changes in idle probability ($\mathscr{P}_I$) and system reliability ($\mathscr{P}_{re}$) remain minimal, likely due to the infrequent nature of repair periods. There's a notable shift from balking to reneging,

with customers choosing to join ($\mathscr{B}_r$ decreases) increases but becoming impatient ($\mathscr{R}_r$ increases) during ongoing repairs. Further, the mean system length ($\mathscr{L}_s$) and expected delay ($\mathscr{E}_{delay}$) show moderate increases, reflecting accumulating queues. In addition, The total expected cost increases, though less dramatically than during vacation periods, indicating the nuanced impact of joining during system maintenance.

4. These findings underscore the delicate balance in system design, particularly for power-saving mechanisms. Joining probabilities significantly influence system performance, with each operational period presenting unique challenges and opportunities for optimization.

4.8 Conclusion and further research

This study presented a queueing analysis of the IEEE 802.16e Power Saving Mechanism using an $M/M/1$ vacation model. Our approach, incorporating differentiated vacation policies, balking and reneging, catastrophes, and working breakdowns, provides a robust framework for understanding the complex dynamics of energy efficiency in wireless networks from a queueing perspective.

The derived steady-state probability distributions and performance measures offer valuable insights into system behavior under various conditions, enabling a deeper understanding of queue dynamics in power-saving scenarios. Our cost model enhances the practical applicability of these findings, offering a comprehensive framework for assessing the economic ramifications of power-saving mechanisms in wireless networks.

While this research significantly advances the field, several avenues for future work remain:

1. Multi-class queueing models: Future studies could extend our model to multi-class queueing systems, representing different types of traffic or priority levels in the network.

2. Non-Markovian arrival and service processes: Exploring non-exponential distributions for arrival and service times could lead to more realistic representations of network traffic patterns.

3. Batch arrival and batch service queueing systems: Investigating the impact of batch arrivals or batch service on power-saving mechanisms presents an interesting direction for future research.

4. Retrial queues: Incorporating retrial phenomena into the model could provide insights into the behavior of blocked or dropped packets in power-saving scenarios.

4. Fluid queue approximations: Developing fluid queue models for large-scale systems could provide efficient methods for analyzing power-saving mechanisms in high-traffic scenarios.

5. Vacation interruption policies: Investigating various vacation interruption policies and their impact on system performance and energy efficiency presents a promising area for future work.

Conclusion

Queuing models provide a rigorous framework for analyzing complex industrial and communication systems. In particular, unreliable queueing models are essential for representing environments subject to breakdowns, working breakdowns, vacations, customer impatience, and catastrophic events.

The objective of this thesis was to develop mathematical representations of such systems and to study their performance under realistic conditions. Using advanced tools such as the Matrix Analytic Method (MAM) and the Matrix Geometric Method (MGM), we constructed models that capture multiple sources of unreliability and demonstrated their stability.

Beyond the theoretical contributions, this work highlights the practical importance of queueing models. By estimating key performance measures and optimizing cost functions, the study provides insights into maintenance planning, customer management, and economic evaluation of unreliable systems.

The results confirm that integrating different disruption features into queueing models leads to more robust and realistic representations compared to traditional approaches. This strengthens the link between mathematical modeling and industrial practice.

Future directions

While this thesis has focused on unreliable queueing models with exponential interarrival and service times, future work can extend these results to more general settings. In particular, studying $G/G/\cdot$ models would allow arrivals and service times to follow arbitrary distributions, providing a more realistic representation of manufacturing and communication systems.

Such extensions would capture variability beyond the exponential assumption and could incorporate heavy-tailed or correlated distributions often observed in practice. The challenge lies in developing analytical methods that remain tractable under these general conditions. Combining matrix-analytic techniques with approximation methods or simulation-based approaches may provide promising avenues for future research.

Bibliography

[1] I. Adan and M. Haviv. The $M/G/1$ queue with vacations and exhaustive service. *European Journal of Operational Research*, 277, 03 2019. doi:https://doi.org/10.1016/j.ejor.2019.03.016.

[2] G. A. Afanasyev. Cost optimization of queueing systems with vacations. *Theory of Probability and Its Applications*, 68(2):308–315, 2023. doi:https://doi.org/10.1137/S0040585X97T99143X.

[3] F. Afroun, D. Aïssani, D. Hamadouche, and M. Boualem. Q-matrix method for the analysis and performance evaluation of unreliable $M/M/1/N$ queueing model. *Mathematical Methods in the Applied Sciences*, 41(18):9152–9163, 2018. doi:https://doi.org/10.1002/mma.5119.

[4] A. Ahuja, A. Jain. Fuzzy analysis of a queueing system featuring an unreliable service provider and geometric arrivals by incorporating constant retrial policy and delayed threshold recovery. *Journal of Ambient Intelligence and Humanized Computing*, 14(6):1–20, 2022. doi:https://doi.org/10.1007/s12652-022-04455-y.

[5] S.I Ammar. 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, 2017. doi:https://doi.org/10.1016/j.joems.2016.09.002.

[6] S.I. Ammar, T. Jiang, and Q. Ye. Transient analysis of impatient customers in an $M/M/1$ disasters queue in random environment. *Engineering Computations*, 37 (6):1945–1965, 2020. doi:https://doi.org/10.1108/EC-06-2019-0274.

[7] S.I. Ammar and P. Rajadurai. Performance analysis of preemptive priority retrial, queueing system with disaster under working breakdown services. *Symmetry*, 11 (3):419, 2019. doi:https://doi.org/10.3390/sym11030419.

[8] S. Assar. Matrix geometric method for $M/M/1$ queueing model with and without breakdown atm machines. *American Journal of Engineering Research (AJER)*, 7: 246–252, 2018.

[9] I. Atencia and P. Moreno. The discrete-time $Geo/Geo/1$ queue with negative customers and disasters. *Comput. Oper. Res.*, 31(9):1537–1545, 2004. doi:https://doi.org/10.1016/S0305-0548(03)00107-2.

[10] K. Baby Saroja and V. Suvitha. Markovian phase-type single working vacation queue with breakdown and standby server: An $M/Hyper-Erlang/1$ queueing system. *Mathematics in Applied Sciences and Engineering*, 5(3):185–275, 2024. doi:https://doi.org/10.5206/mase/17434.

[11] S. Bansal and S. Moses. Reliability characteristics of machining systems and some queueing theory regarding the machine repair problem. *International Journal for Research in Applied Science and Engineering Technology*, 10(12): 1182–1187, 2022. doi:https://doi.org/10.22214/ijraset.2022.48150.

[12] D. T. Barrer. Queuing with impatient customers and indifferent clerks. *Operations Research*, 5(5):644–649, 1957.

[13] A. A. Bouchentouf, M. Boualem, L. Yahiaoui, and H. Ahmad. A multi-station unreliable machine model with working vacation policy and customers' impatience. *Quality Technology & Quantitative Management*, 19(6):766–796, 2022. doi:https://doi.org/10.1080/16843703.2020.1804705.

[14] A. A. Bouchentouf, M. Cherfaoui, and M. Boualem. Performance and economic analysis of a single server feedback queueing model with vacation and impatient customers. *OPSEARCH*, 56(1):300–323, 2019. doi:https://doi.org/10.1007/s12597-019-00357-4.

[15] A. A. Bouchentouf, M. Cherfaoui, and M. Boualem. Analysis and performance evaluation of markovian feedback multi-server queueing model with vacation and impatience. *Amer. J. Math. Management Sci.*, 40(3):261–282, 2021. doi:https://doi.org/10.18080/01966324.2020.1842271.

[16] A. A. Bouchentouf, M. Cherfaoui, and M. Boualem. Mathematical analysis of a markovian multi-server feedback queue with a variant of multiple vacations, balking and reneging. *Discrete and Continuous Models and Applied Computational Science*, 30(1):21–38, 2022. doi:https://doi.org/10.22363/2658-4670-2022-30-1-21-38.

[17] A. A. Bouchentouf, M. Cherfaoui, and M. Boualem. Modeling and simulation of bernoulli feedback queue with general customers' impatience under variant vacation policy. *International Journal of Operational Research*, 46(4):451–480, 2023. doi:https://doi.org/10.1504/IJOR.2023.129959.

[18] A. A. Bouchentouf, A. Guendouzi, and S. Majid. On impatience in $M/M/1/N/DWV$ queue with vacation interruption. *Croatian Operational Research Review*, 11(1):21–37, 2020. doi:https://doi.org/10.17535/crorr.2020.0003.

[19] A. A. Bouchentouf, P. Vijaya Laxmi, and N. R. Yatim. Performance analysis of power saving mechanisms in mobile wireless communication: A queueing model with vacations, impatience, and disruptions, submitted.

[20] A. A. Bouchentouf, L. Yahiaoui, M. Kadi, and S. Majid. Impatient customers in markovian queue with bernoulli feedback and waiting server under variant working vacation policy. *Operations Research and Decisions*, 30(4):5–28, 2020. doi:https://doi.org/10.37190/ord200401.

[21] A. A. Bouchentouf, L. Yahiaoui, and I. Ziad. Modeling and optimizing an AMS with DV policy, waiting servers, impatient customers, and failures: A queueing analysis. *Results in Control and Optimization*, 15:100427, 2024. doi:https://doi.org/10.1016/j.rico.2024.100427.

[22] L. Breuer and D. Baum. *An Introduction to Queueing Theory and Matrix-Analytic Methods*. Springer, 2005. ISBN 978-1-4020-3630-9. doi:https://doi.org/10.1007/1-4020-3631-0.

[23] S. R. Chakravarthy. *Introduction to Matrix-Analytic Methods in Queues 1: Analytical and Simulation Approach - Basics*. 2022. ISBN 9781786307323. doi:https://doi.org/10.1002/9781394165421.

[24] S. R. Chakravarthy, Shruti, and R. Kulshrestha. A queueing model with server breakdowns, repairs, vacations, and backup server. *Operations Research Perspectives*, 7:100131, 2020. doi:https://doi.org/10.1016/j.orp.2019.100131.

[25] X. Chao. A queuing network model with catastrophes and product form solution. *Operations Research Letters*, 18(2):75–79, 1995. doi:https://doi.org/10.1016/0167-6377(95)00029-2.

[26] A. Chen and E. Renshaw. The $M/M/1$ queue with mass exodus and mass arrivals when empty. *Journal of Applied Probability*, 34(1):192–207., 1997. doi:https://doi.org/10.2307/3215186.

[27] A. Chettouf, A. A. Bouchentouf, and M Boualem. A Markovian queueing model for telecommunications support center with breakdowns and vacation periods. *Operations Research Forum*, 5(22), 2024. doi:https://doi.org/10.1007/s43069-024-00295-y.

[28] S. Chew. Continuous-service $M/M/1$ queuing systems. *Appl. Syst.Innov.X.*, 2(2):1, 2019. doi:https://doi.org/10.3390/asi2020016.

[29] J. N. Daigle. *Elementary Continuous-Time Markov Chain-Based Queueing Models*, pages 57–105. Springer US, Boston, MA, 2005. doi:https://doi.org/10.1007/0-387-22859-4_3.

[30] C. K. Deena Merit and M. Haridass. A simulation study on the necessity of working breakdown in a state dependent bulk arrival queue with disaster and optional re-service. *Int. J. Ad Hoc Ubiquitous Comput.*, 41(1):1–15, 2022. doi:https://doi.org/10.1504/IJAHUC.2022.10049033.

[31] N. Dehamnia, M. Boualem, and D. Aïssani. Performance and economic analysis of an unreliable single-server queue with general retrial times and varied customer patience levels. *Hacettepe Journal of Mathematics and Statistics*, 54(2): 1–28, 2025. doi:https://doi.org/10.15672/hujms.1485216.

[32] M. Demircioglu, H. Bruneel, and S. Wittevrongel. Analysis of a discrete-time queueing model with disasters. *Mathematics*, 9(24):32–83, 2021. doi:https://doi.org/110.3390/math9243283.

[33] M. Deruyck, W. Vereecken, W. Tanghe, W. Joseph, M. Pickavet, L. Martens, and P. Demeester. Comparison of power consumption of mobile WiMAX, HSPA and LTE access networks. In *Conf. on Telecommunications Internet and Media Techno Economics (CTTE)*, 2010. Conference proceedings.

[34] A. Di Crescenzo, V. Giorno, A. Nobile, and L.M. Ricciardi. On the M/M/1 queue with catastrophes and its continuous approximation. *Queueing Systems*, 43(4): 329–347, 2003. doi:10.1023/A:1023261830362.

[35] K. Divya and K. Indhira. A literature survey on queueing model with working vacation. *Reliability: Theory & Applications*, 19(1 (77)):40–49, 2024. doi:https://doi.org/10.24412/1932-2321-2024-177-40-49.

[36] B. T. Doshi. Queueing systems with vacations-a survey. *Queueing systems*, 1(1): 29–66, 1986. doi:https://doi.org/10.1007/BF01149327.

[37] S.A. Dudin, O.S. Dudina, and O.I. Kostyukova. Analysis of a queuing system with possibility of waiting customers jockeying between two groups of servers. *Mathematics*, 11(6), 2023. doi:https://doi.org10.3390/math11061475.

[38] D. Fiems. Queues with working vacations: A survey. *Mathematics*, 13:1894, 06 2025. doi:https://doi.org/10.3390/math13111894.

[39] S. C. Graves. The application of queueing theory to continuous perishable inventory systems. *Management Science*, 28(4):400–406, 1982.

[40] G. K. Gupta, A. Banerjee, and U. C. Gupta. On finite-buffer batch-size-dependent bulk service queue with queue-length dependent vacation. *Quality Technology & Quantitative Management*, 17(5):1–27, 2019. doi:https://doi.org/10.1023/A:1023261830362.

[41] P. Gupta and N. Kumar. Performance analysis of retrial queueing model with working vacation, interruption, waiting server, breakdown and repair. *Journal of Scientific Research*, 13(3):833–844, 2021. doi:https://doi.org/10.3329/jsr.v13i3.52546.

[42] F. A. Haight. Queuing with balking I. *Biometrika*, 44(3-4):360–369, 1957.

[43] F. A. Haight. Queueing with reneging. *Metrika*, 2(1):186–197, 1959.

[44] Q.M. He. *Fundamentals of Matrix-Analytic Methods*. Springer New York, 2014. doi:https://doi.org/10.1007/978-1-4614-7330-5.

[45] X. He and W. Hu. Modeling relief demands in an emergency supply chain system under large-scale disasters based on a queuing network. *The Scientific World Journal*, 2014(1):2356–6140, 2014. doi:https://doi.org/10.1155/2014/195053.

[46] M. Houalef, A. A. Bouchentouf, and L. Yahiaoui. 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(1):459–487, 2022. doi:https://doi.org/10.1007/s40305-021-00384-3.

[47] O. Ibe and O. Isijola. $M/M/1$ multiple vacation queueing systems with differentiated vacations. *Modelling and Simulation in Engineering*, 2014(1):1–6, 2014. doi:https://doi.org/10.1155/2014/158247.

[48] O. Isijola and O. Ibe. $M/M/1$ multiple vacation queueing systems with differentiated vacations and vacation interruptions. *IEEE Access*, 2014(1): 1384–1395, 2014. doi:https://doi.org/10.1155/2014/158247.

[49] G. Jain and K. Sigman. A pollaczek–khintchine formula for M/G/1 queues with disasters. *Journal of Applied Probability*, 33(4):1191–1200, 1996. doi:https://doi.org/10.2307/3214996.

[50] J. L. Jain, S. G. Mohanty, and W. Böhm. *A course on queueing models*. CRC Press, 2006.

[51] N. K. Jain and R. Kumar. Transient solution of a catastrophic-cum-restorative queuing problem with correlated arrivals and variable service capacity. *International Journal of Information Management Sciences*, 18(4):461–465, 2007.

[52] J. Janacek and M. Kvet. An approach to uncertainty in emergency service systems via scenarios and fuzzy values. *Croatian Operational Research Review*, 8(1): 237–248, 2017.

[53] B. Jeyakumar, S. Logapriya. Production analysis of manufacturing industry in a single vacation policy under disaster. *Mathematical Models in Engineering*, 9(4): 198–207, 2023. doi:https://doi.org/10.21595/mme.2023.23737.

[54] T. Jiang and L. Liu. 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, 2017. doi:http://dx.doi.org/10.1080/00207160.2015.1128531.

[55] T. Jiang, L. Liu, and J. Li. Analysis of the $M/G/1$ queue in multi-phase random environment with disasters. *Journal of Mathematical Analysis and Applications*, 430(2):857–873, 2015. doi:https://doi.org/10.1016/j.jmaa.2015.05.028.

[56] F. Jolai, S. M. Asadzadeh, and M. R. Taghizadeh. Performance estimation of an email contact center by a finite source discrete time $Geo/Geo/1$ queue with disasters. *Computers & Industrial Engineering*, 55(3):543–556, 2008. doi:https://doi.org/10.1016/j.cie.2008.01.009.

[57] K. Kalidass and R. Kasturi. A queue with working brekdowns. *Comput. Ind. Eng.*, 63(4):719–783, 2012. doi:https://doi.org/10.1016/j.cie.2012.04.018.

[58] V. Karthick and V. Suvitha. Time dependent behavior of a single server queuing system with diffreniated working vacations sunject to system disaster. 18(4): 1019–1031, 2023. doi:https://doi.org/10.24412/1932-2321-2023-476-1019-1031.

[59] J. C. Ke, C. Wu, and G. Zhang. Recent developments in vacation queueing models : A short survey. *International Journal of Operations Research*, 7(4):3–8, 2010.

[60] B. K. Kim and D. H. Lee. The $M/G/1$ queue with disasters and working breakdowns. *Applied Mathematical Modelling*, 38:1788–1798, 2014. doi:https://doi.org/10.1016/j.apm.2013.09.016.

[61] A. Kuki, T. Berczes, J. Sztrik, and A. Toth. Reliability analysis of a retrial queueing systems with collisions, impatient customers, and catastrophic breakdowns. pages 254–259, 2021. doi:https://doi.org/10.1109/IDT52577.2021.9497613.

[62] A. Kumar, M. Boualem, A. A. Bouchentouf, and Savita. *Optimal Analysis of Machine Interference Problem with Random Switching Failure, Vacation Interruption and Synchronized Reneging*, pages 155–168. 2022. ISBN 9781003089636. doi:https://doi.org/10.1201/9781003089636-10.

[63] B. K. Kumar and D. Arivudainambi. Transient solution of an $M/M/1$ queue with catastrophes. *Computers & Mathematics with Applications*, 40(10):1233–1240, 2000. doi:https://doi.org/10.1016/S0898-1221(00)00234-0.

[64] P. Kumar and M. Jain. Reliability analysis of a multi-component machining system with service interruption, imperfect coverage, and reboot. *Reliability Engineering & System Safety*, 202(106):106991, 2020. doi:https://doi.org/10.1016/j.ress.2020.106991.

[65] G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling*. ASA-SIAM Series on Statistics and Applied Probability. Society for Industrial and Applied Mathematics, 1999. ISBN 9780898719734.

[66] D. H. Lee. Optimal pricing strategies and customers' equilibrium behavior in an unobservable $M/M/1$ queueing system with negative customers and repair. *Mathematical Problems in Engineering*, 2017(2):1–11, 2017. doi:https://doi.org/10.1155/2017/8910819.

[67] D. H. Lee, W. S. Yang, and H. M. Park. $Geo/G/1$ queues with disasters and general repair times. *Appl. Math. Model.*, 35(4):1561–1570, 2011. doi:https://doi.org/10.1016/j.apm.2013.05.012.

[68] Y. Levy and U. Yechiali. Utilization of idle time in an$M/G/1$ queueing system. *Management Science*, 22(2):202–211, 1975. doi:https://doi.org/10.1287/mnsc.22.2.202.

[69] X. Li, J. Wang, and F. Zhang. New results on equilibrium balking strategies in the single-server queue with breakdowns and repairs. *Applied Mathematics and Computation*, 241:380–388, 2014. doi:https://doi.org/10.1016/j.amc.2014.05.025.

[70] S. Majid and P. Manoharan. Analysis of an $M/M/c$ queue with single and multiple synchronous working vacations. *Applications and Applied Mathematics: An International Journal (AAM)*, 12(2):671–694, 2017.

[71] S. Majid and P. Manoharan. Analysis of an $M/M/1$ queue with working vacation and vacation interruption. *Applications and Applied Mathematics: An International Journal (AAM)*, 14(1):19–33, 2019. doi:https://doi.org/10.15177/aam.5979.

[72] S. Muthukumar and J. Ebenesar Anna Bagyam. A steady-state behaviour of an $M/M/1$ queue with optional differentiated working vacations, server breakdown, and customer balking. *Advances and Applications in Statistics*, 92(4):603–631, 02 2025. doi:https://doi.org/10.17654/0972361725025.

[73] N. Nandy and S. Pradhan. 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, 2021. doi:https://doi.org/10.1080/16843703.2021.1882655.

[74] N. Nandy and S. Pradhan. 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, 2023. doi:https://doi.org/10.1080/03610926.2021.1966469.

[75] R. Nelson. *Probability, Stochastic Processes, and Queueing Theory: The Mathematics of Computer Performance Modeling*. Springer-Verlag, 1995.

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

[77] R. Padmavathy, K. Kalidass, and K. Ramanath. Vacation queues with impatient customers and a waiting server. *Int. Jour. of Latest Trends in Soft. Eng.*, 1:10–19, 2011.

[78] A. Panta, R.P Ghimire, Dinesh Panthi, and Siddharaj Pant. A review of vacation queueing models in different framework. *Annals of Pure and Applied Mathematics*, 24(2):99–121, 2021. doi:https://doi.org/10.22457/apam.v24n2a02849.

[79] H. M. Park, W. S. Yang, and K. C. Chae. Analysis of the $GI/Geo/1$ queue with disasters. *Stochastic Analysis and Applications*, 28(1):44–53, 2009. doi:https://doi.org/10.1080/07362990903417938.

[80] P. Lidiya and K. Julia Rose Mary. A study on the performance of a queuing system with heterogeneous arrivals and various types of breakdowns under multiple working vacations. *Oper. Res. Decis.*, 34(4):125–140, 2024. doi:https://doi.org/10.37190/ord240408.

[81] M. Pickavet, W. Vereecken, S. Demeyer, P. Audenaert, B. Vermeulen, C. Develder, D. Colle, D. Dhoedt, and P. Demeester. Worldwide energy needs for ict: The rise of power-aware networking. In *2008 IEEE ANTS Conference*, Bombay, India, 2008.

[82] K. Banu Priya, M. Sandeep Kumar, M. Geetha, J. Prabhu, V. Maheshwari, Sukumar Rajendran, M. Prasanna, and P. Rajendran. Queueing network model with jockeying to reduce the waiting time in the airport. *International Journal of System of Systems Engineering*, 11(3-4):363–379, 2021. doi:https://doi.org/10.1504/IJSSE.2021.121471.

[83] R. Raj and V. Jain. Optimization of traffic control in $MMAP[k]/PH[k]/S$ catastrophic queueing model with ph retrial times and preemptive repeat policy. *arXiv preprint arXiv:2203.02508*, 2022. doi:https://doi.org/10.48550/arXiv.2201.08182.

[84] H. Ramdani, A. A. Bouchentouf, and L. Yahiaoui. Optimization analysis of multi-server repair system with bernoulli schedule working vacation, threshold-based recovery policy, and impatience. *Reliability:Theory & Applications*, 20(1):981–995, 2025. doi:https://doi.org/10.24412/1932-2321-2025-182-981-995.

[85] H. Ramdani, A. A. Bouchentouf, and L. Yahiaoui. Finite-capacity multi-server queues with catastrophes, repairs, bernoulli feedback, and fluid approximations. *International Journal of Operational Research*, 1, 01 2025. doi:https://doi.org/10.1504/IJOR.2025.10074923.

[86] K. Ramya, P. Senniappan, and C. Vijayalakshmi. Design and analysis of a vacation model for two-phase queueing system with gated service. *Procedia Computer Science*, 50, 12 2015. doi:https://doi.org/10.1016/j.procs.2015.04.044.

[87] R. Rashid, S. F. Hoseini, M. R. Gholamian, et al. Application of queuing theory in production-inventory optimization. *J Ind Eng Int*, 11(1):485–494, 2015. doi:https://doi.org/10.1007/s40092-015-0099-5.

[88] A. Riska. *Aggregate matrix-analytic techniques and their applications*. Dissertation, The College of William & Mary in Virginia, 2002.

[89] M. Selvaraj, E. Anna Bagyam, and K. Basarikodi. Performance optimization of $M/M/1$ queues with working vacations and server breakdowns in wireless communication systems. *Journal of Telecommunications and Information Technology*, pages 79–85, 2025. doi:https://doi.org/10.26636/jtit.2025.3.2195.

[90] L. D. Servi and S. G. Finn. $M/M/1$ queues with working vacations. *Performance Evaluation*, 50(1):41–52, 2002. doi:https://doi.org/10.1016/S0166-5316(02)00013-5.

[91] M, Shanmugavalli, k.Ignatia Majella, N,Vijaya, G. Senthil, S. Thalapathiraj, and M, Mercy. Analyzing $M/M/1$ working vacation model with dual server breakdowns, retrial queue, and interruptions: A study on classical retrial policies. *International Journal of Advanced Science and Engineering*, 11(4):4492–4504, 2025. doi:https://doi.org/10.29294/IJASE.11.4.2025.4492-4504.

[92] S. Sharma, R. Kumar, B.S. Soodan, and P. Singh. Queuing models with customers' impatience: A survey. *International Journal of Mathematics in Operational Research*, 2022. doi:https://doi.org/10.1504/IJMOR.2022.10052442.

[93] Y. W. Shin. Multi-server retrial queue with negative customers and disasters. *Queueing Systems*, 55(4):223–237, 2007. doi:https://doi.org/10.1007/s11134-007-9018-9.

[94] W. J. Stewart. *Probability, Markov Chains, Queues, and Simulation: The Mathematical Basis of Performance Modeling*. Princeton University Press, Princeton, New Jersey, 2009. ISBN 978-0-691-14062-9.

[95] R. Sudesh and R. S. Priya. An analysis of discrete-time $Geo/Geo/1$ queue with feedback, repair and disaster. *International Journal of Operations Research*, 35(1): 37–53, 2019. doi:https://doi.org/10.1504/IJOR.2019.099542.

[96] W. Sun, S. Li, Y. Wang, and N. Tian. Comparisons of exhaustive and nonexhaustive $M/M/1/N$ queues with working vacation and threshold policy. *Journal of Systems Science and Systems Engineering*, 28, 03 2019. doi:https://doi.org/10.1007/s11518-018-5400-x.

[97] Y. Sun, Z. Ma, and T. Xu. A gated service single vacation $M/G/1$ queue system with setup and closedown times and different customer arrival rates. *Acta Mathematica Scientia*, 46:407–426, 12 2025. doi:https://doi.org/10.1007/s10473-026-0122-5.

[98] M. I. G. Suranga Sampath and L. Jicheng. Impact of customers' impatience on an $M/M/1$ queueing system subject to differentiated vacations with a waiting server. *Qual. Technol. Quant. Manag.*, 17(2):125–148, 2020. doi:https://doi.org/10.1080/16843703.2018.1555877.

[99] v. Suvitha and R. Ohapriyadharsini. Single server markovian queueing system with working vacation, impatient customers and disasters. *Mathematics in Applied Sciences and Engineering*, 6(1):1–93, 2025. doi:https://doi.org/10.5206/mase/21096.

[100] H. Takagi. *Queueing Analysis: Vacation and priority systems (pt. 1)*. Queueing Analysis: A Foundation of Performance Evaluation. North-Holland, 1991. ISBN 9780444889102.

[101] N. Tian and Z.G. Zhang. *Vacation Queueing Models: Theory and Applications*. International Series in Operations Research & Management Science. Springer US, 2006. URL https://books.google.dz/books?id=qX5CAAAAQBAJ.

[102] N. Tian, Q. L. Li, and J. Gao. Conditional stochastic decompositions in the $M/M/c$ queue with server vacations. *Commun.Stat.-Theory Methods.Stoch. Models.*, 15(2):367–377, 1999. doi:https://doi.org/10.1016/j.orl.2006.12.007.

[103] N. Tian, J. Li, and G. Zhang. Matrix analytic method and working vacation queues - a survey. *International Journal of Information and Management Sciences*, 20(4): 603–633, 2009.

[104] D. Towsley and S. K. Tripathi. A single server priority queue with server failures and queue flushing. *Operations Research Letters*, 10(6):353–362, 1991. doi:https://doi.org/10.1016/0167-6377(91)90008-D.

[105] M. Vadivukarasi, K. Kalidass, and R. Jayaraman. Discussion on the optimization of finite buffer markovian queue with differentiated vacations. In *Soft Computing: Theories and Applications*, pages 523–534. Springer, Singapore, 2022.

[106] S. Varshney, M. Bajaj, K. K. Choudhary, M. Pushkarna, and I. Zaitsev. Performance analysis and reliability prediction of multi-state service systems with multiple failure modes of unreliable server: An engineering perspective. *Engineering Reports*, 7(7):e70268, 2025. doi:https://doi.org/10.1002/eng2.70268.

[107] P. Vijaya Laxmi, P. Rajesh, and T. W. Kassahun. Analysis of a variant working vacation queue with customer impatience and server breakdowns. *International*

Journal of Operational Research, 40(4):437–459, 2021. doi:https://doi.org/10.1080/17509653.2019.1653237.

[108] J. Wang, S. Cui, and Z. Wang. Equilibrium strategies in $M/M/1$ priority queues with balking. *Production and Operations Management*, 28(1):43–62, 2018. doi:https://doi.org/10.1111/poms.12889.

[109] D. Wu and H. Takagi. $M/G/1$ queue with multiple working vacations. *Performance Evaluation*, 63(7):654–681, 2006. doi:https://doi.org/10.1016/j.peva.2005.05.005.

[110] W.Xu, L. Liu, L.Li, Z.Wang, and S.Wittevrongel. Analysis of a collision-affected $M/GI/1/N$ retrial queuing system considering negative customers and transmission errors. *Mathematics*, 11(16):1–23, August 2023. doi:https://doi.org/10.3390/math11163532.

[111] W. S. Yang and K. C. Chae. A note on $GI/M/1$ queue with poisson negative arrival. *J. Appl. Prob.*, 38(4):1081–1085, 2001.

[112] Xi.Yang, Y.Zhang, B.Wang, and X.Jun. Li. Performance analysis and cost optimization of the $M/M/1/N$ queueing system with working vacation and working breakdown. *Mathematics*, 13(18), 2025. doi:https://doi.org/10.3390/math13182980.

[113] N.R. Yatim, A. A. Bouchentouf, and P. Vijaya Laxmi. A queueing model for an automatic manufacturing system with disasters, breakdowns and vacations. optimal design and analysis. *Operations Research and Decisions*, 34(4):231 – 250, 2024. doi:https://doi.org/10.37190/ord240413.

[114] D. Yue and W. Yue. Analysis of an $M/M/c/N$ queueing system with balking, reneging, and synchronous vacations. In *Advances in Queueing Theory and Network Applications*, pages 165–180. Verlag, Springer, New York, 2009.

[115] D. Yue, W. Yue, and Y.P. Sun. Performance analysis of an $M/M/c/N$ queueing system with balking, reneging and synchronous vacations of partial servers. pages 128–143. Proc. 6th International Symposium on Operations Research and Its Applications, 2006.

[116] M. Zaghouani, J. Sztrik, and A. Uka. Simulation of the performance of cognitiveradio networks with unreliable servers. *Annales Mathematicae et Informaticae*, 51:1–11, 2020. doi:https://doi.org/10.33039/ami.2020.01.002.

[117] M. Zhang and S. Gao. The disasters queue with working breakdowns and impatient customers. *RAIRO - Operations Research*, 54(3):815 – 825, 2020. doi:https://doi.org/10.1051/ro/2019036.

[118] I. Ziad, P. Vijaya Laxmi, E. G. Bhavant, A. A. Bouchentouf, and S. Majid. A matrix geometric solution of a multi-server queue with waiting servers and customers' impatience under variant working, vacation and vacation interruption. *Yugosl. J. Oper. Res.*, 33(3):389–407, 2023. doi:https://doi.org/10.2298/YJOR220315001Z.