KASDI MERBAH UNIVERSITY OUARGLA

Faculty of Mathematics and Material Sciences

N° d'ordre :
N° de série :

Department of Mathematics

DOCTORAL DISSERTATION

Specialty: Mathematics

Option: Probabilies and Statistics

Present by: Hayat Ramdani

Theme

ANALYSIS OF A MARKOVIAN QUEUE WITH SERVER BREAKDOWNS

Represented in:15/05/2025

Jury members:

Chairman:	Mabrouk MEFLAH	Prof.	UKM , Ouargla
Supervisor:	Amina Angelika BOUCHENTOUF	Prof.	UDL, Sidi Bel Abbes
Co-Supervisor:	Lahcene YAHIAOUI	A. Prof.	UDMT, Saida
Examiner:	Abdelrazzak GHAZAL	Prof.	UKM , Ouargla
Examiner:	Moustapha Saouli	A. Prof.	UKM , Ouargla
Examiner:	Latifa Debbi	Prof.	ENP, Alger

University year: 2024/2025

Dedication

This thesis is dedicated to my parents.

Acknowledgment

In the name of ALLAH, the Most Gracious and the Most Merciful Alhamdulillah, all praises to ALLAH for the strengths and His blessing in completing this thesis. I wish to thank all those who, in one way or another, helped me in the realisation of this work. I am immensely grateful to my supervisor, Professor Amina Angelika Bouchantouf for her invaluable help, advice, suggestions, She has been very supportive and patient throughout the progress of my thesis. Her inspiring comments and suggestions motivated me to continue my research work. I received a great deal of experience and information from her amazing guidance and support during my PhD studies that will be useful to me going forward. Without her support, I never would have been able to accomplish the aim.

I am especially grateful to Doctor Yahiaoui Lahcene for assisting me with typing this work and for his unending support and encouragement throughout the completion of my thesis. I also am indebted to Professor Mefleh Mabrouk and Doctor Abdelmalik Keddi for theirs help in searching for a supervisor to complete my thesis.

ملخص

هذه الأطروحة مخصصة للدراسة المتعمقة لمختلف أنظمة الانتظار الماركوفية متعددة الخوادم المعرضة للخطأ، مما يعكس الديناميكيات المعقدة الكامنة في عمليات الاتصالات الحديثة والأنظمة الصناعية.

تتمحور الورقة حول ثلاث دراسات مترابطة: أولاً، ننظر إلى تحليل نظام الانتظار ذو السعة اللانهائية، والذي يتميز بوصول الدفعات، وردود الفعل برنولي، وانهيار النظام، ونفاد صبر العملاء (الذي يتجلى في الرفض والتراجع). نحن نحدد شروط الاستقرار ونستمد حلول الحالة المستقرة باستخدام وظائف توليد الاحتمالات. ويخضع هذا النموذج لتقييم مقاييس الأداء، فضلا عن تحليل مفصل للتكلفة والعائد.

تركز الدراسة الثانية على نظام الانتظار ذو السعة المحدودة، ودمج حالات اعطال المجدولة، والاسترداد القائم على العتبة، والعطلات النشطة، وانقطاع عطلة برنولي القانونية، ونفاد صبر العملاء، والاحتفاظ بالعملاء غير الصبر. يتم إجراء تحليل الحالة المستقرة باستخدام طريقة مصفوفة Q ، حيث نقوم باستخلاص مؤشرات الأداء الرئيسية وتحسين دالة التكلفة من خلال البحث المباشر وأساليب شبه نيوتن. يتيح لنا هذا الأسلوب تحديد العدد الأمثل للخوادم، والسعة المثلى للنظام، ومعدلات الخدمة المثلى أثناء فترات الانقطاع والتشغيل العادي. أخيرًا، يركز الجزء الثالث من عملنا على تحليل نموذج قائمة الانتظار $M/M/c/N$ مع ردود فعل برنولي والكوارث والإصلاحات. نحن نقدم حلولاً للحالة العابرة و الحالة المستقرة، مُكمّلةً بتحليل سلس لقائمة الانتظار. تهدف الأساليب المنهجية المستخدمة في هذه الأطروحة إلى سد الفجوة بين نماذج الطابور النظرية وتطبيقاتها العملية. ومن ثم فإنها تمهد الطريق لتصميم وإدارة أكثر كفاءة للأنظمة في مختلف المجالات، مثل الاتصالات السلكية واللاسلكية والتصنيع وصناعات الخدمات.

الكلمات المفتاحية أنظمة الطوابير، الأعطال، عدم صبر العملاء، نموذج التكلفة.

Abstract

This thesis is dedicated to the comprehensive study of various multi-server Markovian queueing systems subject to breakdowns, thus reflecting the complex dynamics inherent in the operations of modern communication and industrial systems.
The document is structured around three interconnected studies:
Initially, we focus on analyzing an infinite-capacity queueing system characterized by batch arrivals, Bernoulli feedback, disaster, working breakdowns, and customer impatience (through balking and reneging). We establish stability conditions and derive steady-state solutions using probability generating functions. This model undergoes an evaluation of performance measures, as well as a detailed cost-benefit analysis.
The second study examines a finite-capacity queueing system, incorporating server breakdowns, threshold-based recovery, working vacations, Bernoulli-schedule vacation interruption, customer impatience, and retention of impatient customers. The steady-state analysis is conducted using the Q matrix method. We derive key performance indicators and optimize the cost function through direct search and quasi-Newton methods. This approach allows us to determine the optimal number of servers, the optimal system capacity, as well as the optimal service rates during breakdown and normal operating periods.
Finally, the third part of our work focuses on analyzing an $M/M/c/N$ queueing model with Bernoulli feedback, catastrophes, and repairs. We provide both transient and steady-state solutions, complemented by a fluid queue analysis.
The methodological approaches deployed in this thesis aim to bridge the gap between theoretical queueing models and their practical applications. They thus pave the way for more efficient system design and management in various fields, such as telecommunications, manufacturing, and service industries.
Keywords: Queueing systems, breakdowns, customers' impatience, cost model.

Résumé

Cette thèse se consacre à l'étude approfondie de divers systèmes de files d'attente Markoviens multi-serveurs sujets à des pannes, reflétant ainsi la dynamique complexe inhérente aux opérations des systèmes de communication et industriels modernes.

Le document s'articule autour de trois études interconnectées: Dans un premier temps, nous nous penchons sur l'analyse d'un système de files d'attente à capacité infinie, caractérisé par des arrivées par lots, Bernoulli feedback, catastrophe, des pannes de travail et l'impatience des clients (manifestée par le balking et le reneging) Nous établissons les conditions de stabilité et dérivons les solutions à l'état stationnaire à l'aide de fonctions génératrices de probabilités. Ce modèle fait l'objet d'une évaluation des mesures de performance, ainsi que d'une analyse coût-bénéfice détaillée.

La deuxième étude porte sur un système de files d'attente à capacité finie, intégrant des pannes de serveur, une récupération basée sur un seuil, des vacances actives, une interruption des vacances selon la loi de Bernoulli, l'impatience des clients et la rétention des clients impatients. L'analyse à l'état stationnaire est effectuée en utilisant la méthode de la matrice Q. Nous dérivons des indicateurs de performance clés et optimisons la fonction de coût par le biais de méthodes de recherche directe et quasi-Newton. Cette approche nous permet de déterminer le nombre optimal de serveurs, la capacité optimale du système, ainsi que les taux de service optimaux pendant les périodes de pannes et de fonctionnement normal.

Enfin, la troisième partie de notre travail se concentre sur l'analyse d'un modèle de files d'attente $M/M/c/N$ avec Bernoulli feedback, catastrophes et réparations. Nous proposons des solutions tant transitoires que stationnaires, complétées par une analyse de files d'attente fluides.

Les approches méthodologiques déployées dans cette thèse visent à combler l'écart entre les modèles de files d'attente théoriques et leurs applications pratiques. Elles ouvrent ainsi la voie à une conception et une gestion plus efficaces des systèmes dans divers domaines, tels que les télécommunications, la fabrication et les industries de services.

Mots clés: Systèmes de files d'attente, pannes, impatience des clients, modèle de coût.

List of works

A lot of research has been done, as well as oral and poster presentations, in order to prepare this PhD thesis.

List of research works

1. Ramdani, H., Bouchentouf, A. A., and Yahiaoui, L. (2024). A Markovian batch arrival queueing system with disasters, working breakdowns, and impatience: Mathematical modeling and economic analysis. Yugoslav Journal of Operations Research, http://dx.doi.org/10.2298/YJOR230515037R.

2. Ramdani, H., Bouchentouf, A. A., and Yahiaoui, L. (2024). Finite-capacity multi-server queues with catastrophes, repairs, Bernoulli feedback, and fluid approximations, accepted in International Journal of Operational Research.

3. Ramdani, H., Bouchentouf, A. A., and Yahiaoui, L. (2025). Optimization analysis of multi-server repair system with Bernoulli schedule working vacation, threshold-based recovery policy, and impatience, Reliability: Theory & Applications, 1 (82), 981-995

Presentations

1. H. Ramdani, A. A. Bouchentouf, L. Yahiaoui and A. Rabhi. Retrial queueing model with Bernoulli feedback and abandoned customers. Oral communication at the first International Conference on Pure and Applied Mathematics, IC-PAM'21 May 26-27, 2021, University of Ouargla, Algeria.

2. H. Ramdani, A. A. Bouchentouf and L. Yahiaoui. On feedback queueing system with customers' impatience, multiple working vacations and Bernoulli schedule vacation interruption. Oral communication (online) at 52nd Annual Iranian Mathematics Conference, 30 August–2 September, 2021, Shahid Bahonar University of Kerman, Iran.

3. H. Ramdani, A. A. Bouchentouf and A. Guendouzi. On Markovian queueing model with vacations, waiting server and impatient customers. Oral communication at the National Conference "New Trends in Theoretical and Computational Mathematics", November 8 to 9, 2022, Amine Okkal Hadj Moussa Eg Akhamouk, University of Tamanghasset.

4. H. Ramdani, A. A. Bouchentouf and L. Yahiaoui. On multi-server queue with batch arrival,disasters, and customers' impatience. Oral communication at the Second National Conference on Applied Mathematics and Didactics SNCAMD2023, May 6th, 2023, Constantine, Algeria.

5. H. Ramdani, A. A. Bouchentouf and L. Yahiaoui. On queueing system with batch arrivals and disasters. Oral communication at the International Conference of Young Mathematicians June 1–3, 2023, Institute of Mathematics of NAS of Ukraine (online), Kyiv, Ukraine

Contents

Introdcution

The reliability and efficiency of machining systems are paramount in modern industrial applications, spanning manufacturing, processing, and assembly sectors. These systems operate in environments fraught with uncertainties and potential disruptions, from unexpected breakdowns to customer impatience. As such, the modeling and analysis of machining systems under these challenging conditions have become critical areas of research in operations management and industrial engineering.

This thesis presents a comprehensive exploration of three distinct queueing systems that model various aspects of machining operations under realistic conditions. Through these interconnected research works, we delve into the intricate dynamics of these systems, employing advanced analytical techniques to derive important performance measures and optimize system parameters. Our studies encompass both infinite and finite queueing spaces, utilizing a range of methodologies to address the unique challenges posed by each system.

The subsequent chapters detail each of these research works, presenting the methodologies employed, the results obtained, and their implications for both theoretical advancement and practical applications in the field of operations research and industrial engineering. By addressing both infinite and finite queueing spaces and utilizing a range of analytical techniques, this thesis offers a nuanced and comprehensive exploration of complex machining systems, enabling stakeholders to optimize performance, enhance reliability, and improve cost-effectiveness across a wide range of industrial applications.

The first chapter explores the fundamental concepts of stochastic processes and their application to queueing theory. We begin with an overview of stochastic processes, providing the mathematical foundation for understanding random phenomena evolving over time. The chapter then delves into queueing models, focusing on several key variations that reflect real-world complexities. We examine queueing systems with server breakdowns, which model the impact of equipment failures or maintenance in service systems. The concept of customer impatience is then introduced. We also discuss vacation server models, representing systems where servers become temporarily unavailable. Finally, we present various analytical methods for solving these queueing models, enabling quantitative analysis of system per-

formance. Throughout the chapter, we emphasize the practical applications of these models in diverse fields such as telecommunications, manufacturing, and service industries.
This chapter is based on a paper that has been published in Yugoslav Journal of Operations Research, http://dx.doi.org/10.2298/YJOR230515037R.

The second chapter introduces an advanced queueing model that extends the classical infinite-capacity multi-server Markovian queue by incorporating several complex features: batch arrivals, Bernoulli feedback, working breakdowns, balking, and reneging. A key innovation of this model is its consideration of working breakdowns, where service continues at a reduced rate during repair periods, reflecting real-world scenarios such as degraded performance in computer networks or reduced efficiency in manufacturing systems. The research makes several significant contributions to queueing theory. Firstly, it establishes the stability condition for this sophisticated system, ensuring the long-term existence of a steady state. Secondly, it derives the steady-state solution using probability generating functions (PGFs), a powerful analytical tool in stochastic processes. The study then extracts key performance metrics from the steady-state probability distributions. Furthermore, it formulates a cost model to enable economic analysis of the queueing system. To validate the analytical results and investigate the system's behavior, a comprehensive numerical analysis is conducted. This analysis not only confirms the theoretical findings but also explores the sensitivity of performance measures to various system parameters, providing insights into system optimization and trade-offs in practical applications of this advanced queueing model.

The third chapter presents an in-depth analysis of a sophisticated multi-server Markovian queueing system that incorporates several practically significant features: server breakdowns, threshold-based recovery policy, working vacations, Bernoulli interruption schedule, customer impatience, and retention of reneged customers. This research extends the existing literature, which predominantly focuses on single-server models, by addressing the complexities inherent in multi-server environments. The proposed model offers improved flexibility in characterizing intricate stochastic phenomena within multi-server systems, such as those found in manufacturing or service industries.

The analytical approach employs the Q-matrix method, a powerful technique well-suited for studying quasi-birth-and-death (QBD) processes in steady-state. This methodology enables the derivation of steady-state probabilities and various key performance metrics, providing a comprehensive understanding of the system's behavior under different conditions. The theoretical analysis is complemented by extensive numerical illustrations, which serve to validate the analytical results and offer practical insights.

A significant contribution of this work is the development of a cost function that facilitates the optimization of service rates during both working vacation and normal busy periods. This economic analysis extends to determining the optimal number of servers and exploring the implications of threshold-based recovery policies. These findings provide valuable guidance for system managers and decision-makers in regulating the queueing system efficiently and economically. This chapter is based on a paper that has been published in Reliability: Theory & Applications, 1 (82), 981-995

The fourth chapter explores an innovative $M/M/c/N$ queueing system, incorporating Bernoulli feedback, catastrophes, and repairs. The research extends beyond conventional models by deriving both transient and steady-state solutions, offering a comprehensive view of the system's evolution over time. A key focus is the calculation of time-dependent and time-independent expected queue lengths, providing crucial insights into system behavior under various conditions. The study is further enriched by a fluid queue analysis, which complements the discrete model with a continuous approximation, enabling a deeper understanding of the system's average behavior and facilitating performance optimization strategies.

The inherent complexity of multi-server configurations presents a formidable challenge in queueing theory. These systems, characterized by intricate customer-server interactions, demand rigorous analytical approaches. Despite their complexity, multi-server models are indispensable for representing large-scale operations in sectors such as healthcare, telecommunications, and industrial production. The inclusion of Bernoulli feedback mechanisms allows for a nuanced representation of customer return patterns. This probabilistic approach captures the essence of recurring demand, a critical aspect in many service-oriented environments.

By addressing both catastrophic events and subsequent repairs, the model aligns closely with real-world scenarios. It offers valuable insights into system resilience and recovery, applicable to diverse fields ranging from information technology to manufacturing, where sudden disruptions and restoration processes are common. The dual focus on transient and equilibrium states provides a holistic view of system dynamics. Transient analysis sheds light on initial operational phases, crucial for understanding system behavior during startup or after major disruptions. Conversely, steady-state analysis offers long-term performance predictions, essential for strategic planning and resource allocation.

The fluid queue perspective introduces a macroscopic view of system behavior. This approach, complementing the discrete model, enables the identification of broader trends and patterns, particularly useful for large-scale system design and optimization. By synthesizing these elements, this chapter not only advances theoretical understanding but also bridges the gap between abstract queueing models and practical applications. The findings offer valuable guidance for system designers and managers in diverse fields, enabling more informed decision-making in complex, multi-server environments subject to feedback, catastrophes, and repairs. This chapter has been based on paper that has been accepted in International Journal of Operational Research.

The final section of the thesis outlines the key contributions, exploring their significance and potential limitations. It also suggests possible directions for future research.

Collectively, these three distinct studies contribute to the broader understanding of complex machining systems operating under realistic, challenging conditions. By developing and analyzing these sophisticated queueing models – each with its own unique characteristics and spanning both infinite and finite buffer capacities – this thesis aims to provide valuable insights for system designers and managers. The diverse methodologies employed, including probability generating functions, matrix-analytic methods, and Laplace transforms, allow for a comprehensive analysis of these systems under various conditions.

Chapter 1

Advanced queueing models and their applications

1.1 Stochastic processes

1.1.1 Introduction

Stochastic processes are mathematical frameworks for modeling systems that evolve randomly over time or space. They are widely used in physics, finance, biology, and engineering to analyze phenomena such as stock prices, particle motion, population dynamics, and signal processing. To really dig into queueing models, we first need a solid grasp of these processes. In this section, we will introduce the fundamental concepts and definitions. While we have tailored some definitions to better suit our discussion of queueing theory, much of our foundational understanding here draws from the comprehensive work presented in [46].

A stochastic process is characterized by two key components:

- A state space (denoted S), which defines all possible values the process can take (e.g., discrete states like integers or continuous values like real numbers).
- A parameter space (denoted $\mathcal{T}$), typically representing time (discrete or continuous) or spatial coordinates.

Definition 1.1. *A stochastic process is a family of random variables $\{X_t : t \in \mathcal{T}\}$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is the sample space, $\mathcal{F}$ is a σ-algebra of measurable events, and $\mathbb{P}$ is a probability measure.*

Remark 1.2. *All random variables $\{X_t\}$ in a stochastic process must share a common state space S and be defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. This ensures consistency in modeling trajectories (sample paths) of the process.*

A key concept in the study of stochastic processes is stationarity. A process is called strictly stationary if the joint distribution of $(X_{t_1}, X_{t_2}, \ldots, X_{t_n})$ is identical to the joint distribution of $(X_{t_1+\tau}, X_{t_2+\tau}, \ldots, X_{t_n+\tau})$ for all $\tau, n \in \mathbb{N}$. For weak stationarity, only the mean and covariance structure must remain invariant over time.

1.1.2 Markov processes

While general stochastic processes can exhibit intricate dependency structures, Markov processes offer a simplified yet powerful approach. These processes are characterized by the "memoryless" property, where the future state depends solely on the present, independent of the past. Despite this apparent limitation, Markov processes find wide applicability in modeling real-world phenomena, from population dynamics to financial markets.

The mathematical tractability of Markov processes, compared to more general stochastic processes, makes them particularly valuable in analysis and prediction. Moreover, they often serve as effective approximations for more complex systems.

Definition 1.3. *A Markov chain is a stochastic process $\{X_t : t \in \mathcal{T}\}$ with a discrete (countable) state space S, where the parameter space $\mathcal{T}$ is typically either discrete (e.g., $\mathbb{N}$) or continuous (e.g., $\mathbb{R}^+$), and the process satisfies the Markov property:*

$$\mathbb{P}(X_{t_{n+1}} = x_{n+1} \mid X_{t_1} = x_1, \ldots, X_{t_n} = x_n) = \mathbb{P}(X_{t_{n+1}} = x_{n+1} \mid X_{t_n} = x_n)$$

for any times $t_1 < t_2 < \cdots < t_n < t_{n+1}$.

Markov processes are fully characterized by their transition probabilities $P(s, t, x, \mathcal{A})$ and an initial distribution. Here, $P(s, t, x, \mathcal{A})$ represents the probability of transitioning from state x to any state in set $\mathcal{A}$ during the time interval $[s, t]$. For countable state spaces, this probability can be computed as:

$$\mathbb{P}(s, t, x, \mathcal{A}) = \sum_{y \in \mathcal{A}} \mathbb{P}\{X(t) = y | X(s) = x\}.$$

Definition 1.4. *If the transition probability depends only on the difference of s and t, i.e., $\mathbb{P}(s, t, x, \mathcal{A}) = \mathbb{P}(|t - s|, x, \mathcal{A})$, the Markov process is called a (time-)homogeneous Markov process.*

Definition 1.5. *A Markov process is classified as (time-)homogeneous when its transition dynamics remain consistent regardless of the absolute time points considered. This property is mathematically expressed as:*

$$\mathbb{P}(s, t, x, \mathcal{A}) = \mathbb{P}(|t - s|, x, \mathcal{A}) \tag{1.1}$$

Remark 1.6. *In Equation (1.1), the transition probability depends only on the time interval $|t-s|$, not on the specific values of s and t. This characteristic implies that the process's statistical behavior is invariant under time shifts, making its stochastic properties uniform across all temporal reference frames.*

1.1.3 Markov chains

Markov chains are powerful mathematical tools that enable the computation of limiting distributions under reasonably mild conditions. Their versatility makes them invaluable in modeling diverse real-world phenomena across engineering and scientific disciplines. The computational simplicity of Markov chains is achieved through a strategic restriction of the state space.

Definition 1.7. *A Markov chain is defined as a Markov process with a countable (discrete) state space.*

Remark 1.8.

1. *An equivalent characterization of a Markov chain is that of a stochastic process where the probability of transitioning to any future state depends solely on the current state, not on the sequence of events that preceded it. This property, known as the Markov property, distinguishes Markov chains from other types of stochastic processes.*

2. *The classification of Markov chains is based on their parameter space, resulting in either discrete-time or continuous-time variants. Drawing from the definition of Markov processes, a Markov chain is termed (time-)homogeneous if its transition probabilities are time-independent.*

The subsequent sections will focus exclusively on homogeneous Markov chains.

1.1.3.1 Homogeneous Markov chains in discrete time

In the context of discrete state space S and parameter space $\mathcal{T}$, the transition probabilities introduced in Section 1.1.2 can be simplified to $p_{ij} = \mathbb{P}\{X_n = j \mid X_{n-1} = i\}$. These probabilities can be organized into a transition probability matrix for a single stage:

$$\mathbf{P} = (p_{ij})_{i,j\in\mathcal{T}} = \begin{pmatrix} p_{11} & p_{12} & \cdots \\ p_{21} & p_{22} & \cdots \\ \vdots & \vdots & \end{pmatrix}.$$

Remark 1.9. *It's important to note that "matrix" here is used in a broad sense, encompassing the possibility of infinite dimensions.*

The ℓ-step transition probability, denoted by $p_{ij}^{(\ell)}$, represents the probability of transitioning from state i to state j in exactly ℓ steps. This probability is computed using the Chapman-Kolmogorov equations:

$$p_{ij}^{(\ell)} = \sum_{k \in \mathcal{T}} p_{ik}^{(\ell-1)} p_{kj}, \text{ for } \ell \geq 2.$$

In essence, the ℓ-step transition probabilities are recursively defined by the single-step transition probabilities.

The Chapman-Kolmogorov equations reflect the principle that to reach state j from state i in ℓ steps, the process must pass through some intermediate state k after $\ell - 1$ steps. Since k can be any state in the system, we sum over all possibilities. In matrix notation, the system of Chapman Kolmogorov equations is reduced to the power operation: $\mathbf{P}^{(\ell)} = \mathbf{P}^{\ell}$. For an initial distribution vector a, the system's state after ℓ transitions is represented by $a\mathbf{P}^{\ell}$. Thus, a Markov chain is completely characterized by its initial distribution and transition matrix.

In practical systems, we often expect a steady state to emerge after an initial transient phase. Mathematically, this steady state corresponds to a limiting distribution—a long-term behavior independent of the system's starting configuration. For such a distribution to exist, the Markov chain must be irreducible: visualized as a graph, every state must be reachable from every other state via transitions with nonzero probability. This excludes chains that split into disconnected components.

Definition 1.10. *A Markov chain is termed irreducible if all states intercommunicate, i.e., for any distinct states i and j in $\mathcal{S}$:*

$$\exists \ell : p_{ij}^{(\ell)} > 0.$$

If this condition is not met, the chain is considered reducible.

Definition 1.11. *A state is classified as aperiodic if the greatest common divisor of the set*

$$\{\ell : p_{ii}^{((\ell)} > 0\}$$

is 1. *Otherwise, the state is periodic. Within an irreducibility class, all member states share the same period.*

To establish the existence of a steady state, we must ensure that the Markov chain neither diverges to infinity nor becomes trapped within a subset of states. This necessitates a detailed examination of the visitation patterns for each state.

Definition 1.12. *Let $r_i^{(\ell)}$ denote the probability of first returning to state i after exactly m steps:*

$$r_i^{(\ell)} = \mathbb{P}\{X_\ell = i, X_k \neq i \textit{ for } k = 1, 2, ..., \ell - 1 | X_0 = i\}.$$

Define r_i as the probability of eventually returning to state i:

$$r_i = \sum_{\ell=1}^{\infty} r_i^{(\ell)}.$$

State i is classified as recurrent if $r_i = 1$, and transient if $r_i < 1$. For a recurrent state, we define the mean recurrence time as:

$$\ell_i = \sum_{\ell=1}^{\infty} \ell r_i^{(\ell)}.$$

If $\ell_i < \infty$, state i is positive recurrent; otherwise, it's null recurrent. These properties extend to all members of an irreducibility class.

An alternative characterization of recurrence involves the expected number of visits to a state. Interestingly, this approach yields equivalent results from:

Theorem 1.13. *[42] Let $v_i = \#\{n > 0 : X_n = i\}$ represent the number of visits to state i. The following conditions are equivalent:*

$$r_i = 1 \iff \mathbb{P}\{v_i = \infty\} = 1 \iff \mathbb{E}[v_i] = \infty$$

The confluence of these properties leads to the concept of ergodicity:

Definition 1.14. *A Markov chain is termed ergodic if it is irreducible, aperiodic, and positive recurrent.*

As a Markov chain reaches equilibrium, it approaches a steady-state distribution, denoted as $\pi_i = \lim_{n\to\infty} \mathbb{P}\{X_n = i\}$. These π values are commonly known as equilibrium probabilities. When they exist, the probability vector $\pi = (\pi_i)_{i\in\mathcal{S}}$ remains unchanged by the Markov chain's transitions.
Mathematically, this steady-state behavior is described by the balance equations:

$$\pi = \pi\mathbf{P}, \qquad \sum_{i\in\mathcal{S}} \pi_i = 1.$$

Remark 1.15. *These equations capture two essential characteristics of the equilibrium distribution:*

1. *The equation $\pi = \pi\boldsymbol{P}$ reflects the distribution's invariance under Markov transitions.*

2. *The condition $\sum_{i\in\mathcal{S}} \pi_i = 1$ ensures π is a proper probability distribution.*

Now, we can now present key findings about the existence and nature of the equilibrium distribution π from [32] & [125]:

Theorem 1.16. *[46] Given an aperiodic Markov chain in discrete time, the limits* $\pi_i = \lim_{n\to\infty} \mathbb{P}\{X_n = i\}$ *for all* $i \in \mathcal{S}$ *exist. For an irreducible and aperiodic Markov chain the following expression holds*

$$\pi_i = \frac{1}{\ell_i}.$$

These limits are independent of the initial distribution but do not necessarily constitute a probability distribution, because ℓ_i *might become infinite. In case the underlying Markov chain is ergodic, the vector* $\pi = (\pi)_{i\in\mathcal{S}}$ *represents a valid probability distribution.*

1.1.3.2 Homogeneous Markov chains in continuous time

Continuous-time Markov chains require a different analytical approach compared to their discrete counterparts. We'll explore a conventional method that builds upon discrete-time techniques, with slight adjustments to account for the continuous nature of the parameter.

In this context, transition probabilities are expressed as $p_{ij}(s,t) = \mathbb{P}\{X(t) = j|X(s) = i\}$. Time homogeneity allows for the simplification $p_{ij}(s,t) = p_{ij}(0, t-s) = p_{ij}(t-s)$. It's important to note that the concept of a single-step transition probability doesn't apply here, as there's no discrete time unit.

To analyze continuous-time Markov chains, we must employ infinitesimal calculus, which necessitates additional constraints on the transition rate matrix $\mathbf{P}(t)$:

Definition 1.17. *A matrix* $\boldsymbol{P} = (p_{ij})_{i,j\in\mathcal{S}}$ *is considered stochastic if* $\sum_j p_{ij} = 1$ *for all* $i, j \in \mathcal{S}$*, and each column contains at least one non-zero element.*

Definition 1.18. *A transition semigroup on state space* $\mathcal{S}$ *is defined as* $\boldsymbol{P}(t)$ *if:*

1. $\boldsymbol{P}(t)$ *is a stochastic matrix*
2. $\boldsymbol{P}(0) = \boldsymbol{I}$
3. $\boldsymbol{P}(t+s) = \boldsymbol{P}(t)\boldsymbol{P}(s)$

Remark 1.19. *The third condition is analogous to the Chapman-Kolmogorov equations in continuous time.*

We also assume continuity at 0, meaning $\lim_{t\to 0} \mathbf{P}(t) = \mathbf{P}(0) = \mathbf{I}$. This implies:

$$\lim_{t\to 0} p_{ij}(t) = p_{ij}(0)$$

and

$$q_{ij} = \lim_{t \to 0} \frac{p_{ij}(t) - p_{ij}(0)}{t}, \tag{1.2}$$

where $0 \leq q_{ij} < \infty$ for $i \neq j$ and $q_{ii} \leq 0$. In matrix notation, $\mathbf{Q} = q_{(ij)_{i,j \in S}}$ represents the infinitesimal generator. The matrix equivalent to (1.2) is:

$$\mathbf{Q} = \lim_{t \to 0} \frac{\mathbf{P}(t) - \mathbf{P}(0)}{t} = \lim_{t \to 0} \frac{\mathbf{P}(t) - \mathbf{I}}{t}.$$

Using the infinitesimal generator, we can define additional properties for $\mathbf{P}(t)$:

Definition 1.20. *$\boldsymbol{P}(t)$ is stable if -$q_{ii} < \infty$ for all $i \in \mathcal{S}$. It's conservative if -$q_{ii} = \sum\limits_{j \in \mathcal{S}, j \neq i} q_{ij}$ for all $i \in \mathcal{S}$.*

Remark 1.21. *The conservation property ensures that $\sum\limits_{j \in \mathcal{S}} p_{ij}(t) = 1$ for fixed t, indicating that the process preserves all work performed.*

Reformulating the Chapman-Kolmogorov equations and applying limits, we arrive at Kolmogorov's forward differential system:

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

In component form, this system can be expressed as:

$$\frac{d}{dt}p_{ij}(t) = p_{ij}(t)q_{jj} + \sum_{k \in \mathcal{S}, k \neq j} p_{ik}(t)q_{kj}.$$

Similarly, the backward differential system is given by:

$$\frac{d}{dt}p_{ij}(t) = q_{ii}p_{ij}(t) + \sum_{k \in \mathcal{S}, k \neq i} q_{ik}p_{kj}(t).$$

Remark 1.22. *Continuous-time Markov chains inherit concepts such as irreducibility, communication, transience, and recurrence from their discrete-time counterparts. This is because for any $c > 0$, $Y_n = X(t)$ with $t = cn$ describes a discrete-time Markov chain [58]. However, aperiodicity doesn't apply due to the absence of a fixed time unit.*

To determine the steady-state distribution of a continuous-time Markov chain, we begin with $p_j(0) = \mathbb{P}\{X_0 = j\}$ as the initial probability for state j, and define $\mathbf{p} = (p_j(0))_{j \in S}$ as the initial probability vector. Using the forward differential equation and the law of total probability, $p_j(t) = \sum_{i \in \mathcal{S}} p_{ij}(t)p_i(0)$, we get:

$$\frac{d}{dt}p_j(t) = q_{jj}p_j(t) + \sum_{k \in \mathcal{S}, k \neq j} p_k(t)q_{kj}.$$

Assuming irreducibility ensures the existence of limiting probabilities $\lim_{t\to\infty} p_j(t)$. At equilibrium, $\frac{d}{dt}p_j(t) = 0$, leading to:

$$0 = q_{jj}p_j(t) + \sum_{k\in\mathcal{S},k\neq i} p_k(t)q_{kj}.$$

In matrix notation, this becomes:

$$0 = \mathbf{pQ}.$$

This system of equations represents the concept of global balance. To ensure p_j forms a valid probability distribution, we add the constraint:

$$\sum_{j\in\mathcal{S}} p_j = 1.$$

These equations together determine the stationary probabilities, with $\mathbf{p}$ as the stationary probability vector. A similar derivation exists for Kolmogorov's backward differential system.

Having established methods for computing equilibrium distributions, a natural inquiry arises: What conditions ensure the persistence of these solutions? Drawing from our understanding of system stability and conservation principles, we can identify two fundamental criteria from [32] & [125]:

Theorem 1.23. *[46] Given a conservative continous time Markov chain, Kolmogorov's backward differential system is valid. Kolmogorov's forward-differential system applies for a stable Markov chain in continous time.*

Remark 1.24. *It's worth noting that the global balance equations are applicable to both continuous and discrete cases. For discrete Markov chains, we can construct the infinitesimal generator as* $\boldsymbol{Q} = \boldsymbol{P} - I$.

1.2 On queueing models

Queueing theory, pioneered by Agner Krarup Erlang in the early 20th century, is the predominant analytical framework for modeling systems where entities await resource access and service provision. Erlang's seminal 1909 paper, "Probability Theory and Telephone Conversations," established the foundations of queueing theory, introducing the Poisson and exponential distributions as key components. His subsequent work over two decades developed crucial concepts such as statistical equilibrium and state equilibrium equations. Queueing models have since been applied to diverse phenomena, including communication networks, computer systems, manufacturing processes, and traffic flow, making queuing theory an essential tool for analyzing and optimizing resource allocation in various domains [16].

1.2.1 Core components of queueing systems

A queueing system is defined by the following elements:

- Arrival process: Describes how entities (e.g., customers, data packets) enter the system. The Poisson process is widely used due to its memoryless property, where arrivals are random and independent.
- Service process: Governs how entities are served. Service times often follow exponential distributions.
- Queue discipline: Rules for selecting the next entity to serve. Common disciplines include:
  - First-Come-First-Served (FCFS)/First-In-First-Come (FIFO)
  - Last-Come-First-Served (LCFS, e.g., stack management)
  - Processor Sharing (PS, e.g., CPU time allocation)
  - Priority queues (e.g., emergency triage)
- Number of servers: Single-server (e.g., small clinic) vs. multi-server systems (e.g., cloud computing clusters).
- System capacity: Finite buffers (e.g., call centers with hold limits) vs. infinite waiting rooms (e.g., cloud platforms with dynamic scaling).

1.2.2 Kendall's notation

Queueing systems are classified via Kendall's notation $A/S/c/K/m$, where:

- A: Arrival process (M=Poisson, D=deterministic, GI=renewal process).
- S: Service process (same symbols as A).
- c: Number of servers.
- K: System capacity (default: ∞).
- m: Queue discipline (default: FCFS).

Examples:

- $M/M/1$: Single-server queue with Poisson arrivals and exponential service.
- $GI/G/3/10$: Three-server system with general independent arrivals, general service, and finite capacity of 10.

1.2.3 Performance metrics

Key measures include:

- Queue length: Number waiting (e.g., packets in a router buffer).
- Waiting time: Delay before service starts (critical for user satisfaction).
- Loss probability: The probability of entities being dropped when buffers overflow.
- Server utilization: Fraction of time servers are busy (tradeoff between efficiency and congestion).

1.2.4 Markovian queues: Key models

1.2.4.1 $M/M/1$ queue (Figure 1.1)

- Structure: Single server, Poisson arrivals (λ), exponential service (μ), FIFO (First In First Out) discipline.
- Stability: Requires $\rho = \lambda/\mu < 1$, where ρ is the traffic intensity.
- Steady-state: Stationary distribution $\pi_n = (1-\rho)\rho^n$, where π_n is the probability of n customers in the system.

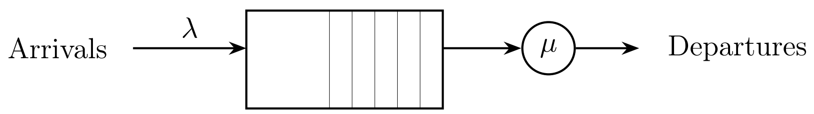

Figure 1.1: Schematic of an $M/M/1$ queue

Figure 1.2 illustrates the state transition diagram.

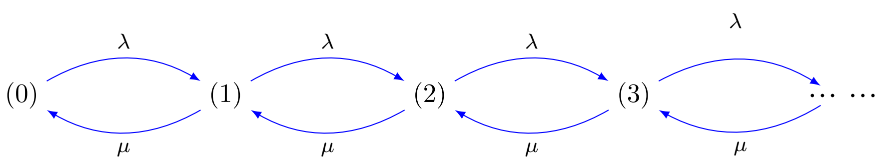

Figure 1.2: transition diagram of an $M/M/1$ queue

1.2.4.2 M/M/c queue (Figure 1.3)

- Structure: c parallel servers (e.g., supermarket checkouts).
- Performance: Reduced waiting times compared to $M/M/1$; stability requires $\lambda < c\mu$.
- Steady-state: For $\rho = \lambda/(c\mu) < 1$,

$$\pi_n = \begin{cases} \frac{(c\rho)^n}{n!}\pi_0, & 0 \le n < c, \\ \frac{c^c\rho^n}{c!}\pi_0, & n \ge c, \end{cases}$$

where

$$\pi_0 = \left[\sum_{k=0}^{c-1} \frac{(c\rho)^k}{k!} + \frac{(c\rho)^c}{c!} \cdot \frac{1}{1-\rho}\right]^{-1}.$$

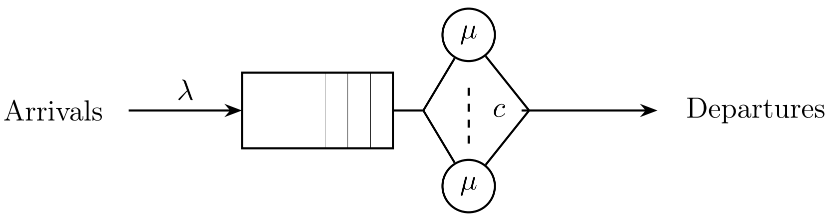

Figure 1.3: The $M/M/c$ model

The transition diagram is illustrated in Figure 1.4.

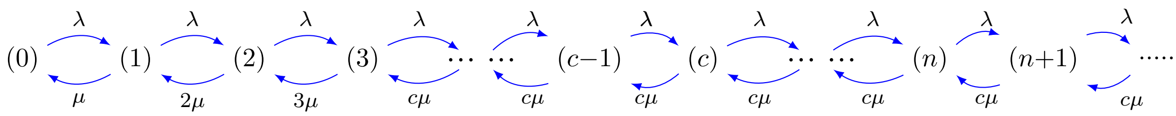

Figure 1.4: Transition diagram of $M/M/c$ queue

1.2.4.3 M/M/∞ queue

- Structure: Infinite servers (e.g., self-service cloud instances).
- Steady-state: Follows a Poisson distribution $\pi_n = e^{-\rho}\rho^n/n!$, where $\rho = \lambda/\mu$.

1.2.5 Extended queueing models

1.2.5.1 Queueing models with breakdowns, working breakdowns, and catastrophes

In real-world queueing systems, servers are subject to various types of disruptions, including breakdowns, working breakdowns, and catastrophes. These events can significantly impact system performance, revenue, and reputation. To mitigate these effects, it is crucial to maintain system reliability at a high level and develop strategies to handle different types of disruptions.

Server disruptions in queueing systems can be classified into three main categories:

1. Breakdowns:
   - Permanent breakdowns: The server becomes permanently inoperable.
   - Temporary breakdowns: The server is temporarily unavailable but returns to service after a repair period.
2. Working breakdowns: The server continues to operate but at a reduced efficiency. This state represents a degraded mode of operation where service is still provided, albeit at a slower rate.
3. Catastrophes: Severe events that affect the entire system, potentially causing simultaneous failures of multiple servers, clearing the queue, or requiring a system-wide reset.

The impact of these disruptions depends on various factors, including the number of servers, customer arrival rate, service rate, and the specific characteristics of each type of disruption. Common causes of server disruptions include:

1. Hardware failures (e.g., CPU, hard drive, memory)
2. Software errors
3. Environmental factors (e.g., temperature, humidity, electromagnetic interference)
4. External events (e.g., power outages, natural disasters)

The consequences of server disruptions in queueing systems may include:

- Increased waiting times
- Reduced system throughput
- Decreased productivity and potential revenue loss

- Diminished customer satisfaction
- In the case of catastrophes, potential system-wide failures or resets

Analyzing the effects of breakdowns, working breakdowns, and catastrophes on queueing systems is essential for developing robust models that accurately reflect real-world scenarios. This analysis can inform strategies for:

- Improving system reliability
- Implementing effective repair and recovery policies
- Optimizing performance under various disruption risks
- Developing contingency plans for catastrophic events

A literature review on queueing models with breakdowns, working breakdowns, and catastrophes reveals a focus on causes, impact, mitigation strategies, and technological advancements in server management. [7] analyzed five types of queueing problems with server breakdown. Subsequent research explored various aspects of these systems, including repairable servers [34], unstable servers [78], and cost analysis of multi-server systems with impatient customers [149]. [57] and [148] investigated models with different types of server breakdowns and optional services. Optimal control strategies for N-policy queues with breakdowns were studied by [151]. [64] examined models with working vacations and multiple breakdown types.

Recent research has focused on more complex scenarios, including working breakdown services [74, 90] and systems with catastrophes [6, 122]. Additional studies have explored impatient customers [126], multiple adapted vacation policies [98], and geometric reneging [45].

1.2.5.2 Customer impatience in queueing systems

The study of queueing systems aims to provide a comprehensive understanding of real-world waiting line scenarios. While traditional queueing models often overlooked customer impatience, this factor is crucial in accurately representing and analyzing congestion issues in both everyday and industrial settings [3].

Customer impatience, manifested as anxiety and frustration during wait times, significantly impacts queueing dynamics. As queue length increases, arriving customers may become discouraged, leading to various behaviors that affect system performance. These behaviors are categorized in queueing theory as:

1. Balking: Customers refuse to join the queue upon arrival. The study of queues with balking has evolved significantly since its inception.

Haight [59] pioneered this field. Since then, many research papers have been done on the subject (cf. Wang et al. [150], Zirem et al. [162], Boualem [19], Wang et al. [86], Hanukov et al. [60])

2. Reneging: Customers leave the queue after joining a line but leaving before being served. The study of queueing systems with reneging has gained significant attention in recent years, with researchers exploring various aspects and applications of this phenomenon. Recent key contributions include Kumar and Sharma [79], Kumar and Sharma [80], Soodan et al. [124], Sudhesh and Azhagappan [128], Dong et al. [47], Ahmadi et al. [1], Bouchentouf et al. [21], and Zhang et al. [160].

3. Jockeying: Customers switch between queues in multi-server systems. The phenomenon of customers switching between queues, has been an important area of study in queueing theory. Important research contributions include Ravid [113], Kumar et al. [112], and Dudin et al. [50].

1.2.5.3 Server vacation models in queueing theory

Server vacation models have emerged as an important area of study, addressing scenarios where servers become temporarily unavailable, often engaged in maintenance tasks or serving secondary customers, to primary customers. These models find applications in various systems, including production systems, and computer and communication systems. Comprehensive reviews and surveys have been provided by Doshi [48], Takagi [135], and Tian and Zhang [62].

Types of vacation policies

1. Multiple vacation policy: The server goes on vacation after serving the customers till queue is empty. Upon his return from vacation, he resumes service if he finds at least one customer; otherwise, he goes for another vacation and repeats the process until he encounters at least one.

2. Single vacation policy: The server takes one vacation after each busy period, returning immediately upon completion.

3. K-Variant Vacation Policy: The server takes up to k consecutive vacations until finding a waiting customer or reaching the maximum number of vacations.

4. Exhaustive service policy: The server serves all waiting customers and new arrivals before taking another vacation.

5. Gated service policy: The server only serves customers present upon return from vacation, taking another vacation after serving this group.

6. Limited service policy: The server serves a maximum number of customers before taking another vacation. A special case is the single-service scheme, where only one customer is served.

The intersection of server vacations and customer impatience in queueing systems has garnered significant attention in recent years. Key contributions include Yechiali and Altman [3], Kalidass et al. [72], Goswami and Selvaraju [55], Panda et al. [106], Ammar [5], Bouchentouf et al. [23], and Goswami and Mund [54], and Ammar et al. [6].

1.2.6 Applications of advanced queueing models

In what follows, we present different applications that demonstrate the versatility and practical relevance of advanced queueing models across various industries, highlighting the importance of continued research and development in this field:

- Telecommunications: In network traffic management, queueing models with breakdowns are crucial for modeling router failures or maintenance. Customer impatience models apply to call centers, where callers may hang up if kept waiting too long. Vacation server models are relevant in modeling sleep modes in energy-efficient networks.

- Healthcare systems: Emergency room patient flow can be modeled using queueing systems with priority classes and customer impatience. Hospital bed allocation benefits from models with server breakdowns, representing equipment failures or staff shortages. Vaccine distribution logistics often involve vacation server models to account for supply chain interruptions.

- Manufacturing and production: Assembly line optimization frequently uses queueing models with breakdowns to account for machine failures. Inventory management systems benefit from vacation server models, representing periods of no production. Quality control processes often incorporate customer impatience models to represent time-sensitive materials.

- Transportation and logistics: Traffic flow modeling employs queueing models with breakdowns to represent accidents or road closures. Airport security checkpoint management uses customer impatience models to optimize throughput. Shipping systems often use vacation server models to represent scheduled maintenance of vehicles.

- Computer systems: CPU task scheduling can be optimized using queueing models with priorities and customer impatience. Database query processing often employs models with server breakdowns to account for system crashes. Cloud computing resource allocation frequently uses vacation server models to represent dynamic scaling of resources.

- Retail and service industries: Supermarket checkout optimization benefits from queueing models with customer impatience. Bank teller systems often use models with server breakdowns to account for technical issues. Restaurant seating and order processing can be improved using vacation server models to represent staff breaks.

- Energy systems: Power grid load balancing often employs queueing models with breakdowns to represent equipment failures. Renewable energy integration benefits from models with vacation servers, representing intermittent availability. Electric vehicle charging station management uses customer impatience models to optimize charging schedules.

1.2.7 Resolution methods for queueing models

1.2.7.1 Method of supplementary variables (SVM)

The method of supplementary variables (SVM) is an analytical technique that transforms non-Markovian queueing systems into Markovian ones by expanding the state space. It is particularly useful for analyzing systems with general service time distributions or complex features that resist straightforward Markovian analysis.

SVM introduces additional variables to capture non-Markovian elements, typically representing residual or elapsed times of certain processes. This approach enables the analysis of a wide range of non-Markovian queueing systems and allows for both transient and steady-state solution derivations.

Key applications of SVM include calculating stationary distributions in systems with general service time distributions, analyzing non-Markovian queues using Markovian techniques, and examining complex queueing systems.

Cox [43] first introduced SVM, laying the foundation for subsequent research. Notable contributions have come from Neuts [102], Baccelli and Bremaud [12], and more recently, Seeniraj and Sundaramoorthy [118] and Kumar and Arumuganathan [92], demonstrating SVM's continued relevance in addressing complex queueing scenarios.

1.2.7.2 Recursive method

Recursive techniques in queueing theory solve complex problems by partitioning them into smaller, manageable subproblems that can be resolved using the same strategy. The original problem is then solved by recursively applying solutions to these subproblems. It is well noted that while recursive methods can be effective for resolving queueing issues, their complexity often leads to the use of iterative approaches in practical applications.

This technique has been widely applied in queueing literature, particularly in the analysis of vacation queueing models. For more recent research works include Bouchentouf et al. [25], Lee et al. [84], Kumar et al. [95], Ahuja [2], Sahana and Baburaj [117], Ma and Guo [91].

1.2.7.3 Matrix analytic methods

Matrix analytic method is a powerful class of techniques in probability theory used to analyze Markov chains with recurring structures. It employs matrix polynomials to describe transition probability matrices, which are then examined using matrix analysis tools. The method is particularly effective for quasi-birth-and-death (QBD) processes, where transitions are limited to neighboring levels, allowing for representation as block matrices.

This techniques enables the calculation of various performance metrics and facilitates the analysis of transient behaviors in QBD processes. It have founds wide application in inventory theory, telecommunications, and queueing theory, proving invaluable for examining Markov chains with repeating structures.

The matrix analytic methodology became a cornerstone in constructing and evaluating queue models. Zhang et al. [138] provided a comprehensive survey of matrix analytic methods in working vacation queues, highlighting the efficacy of Neuts' [101] approach. This survey serves as a foundational resource in understanding the method's applications and developments. Recent advancements have expanded the scope and sophistication of these methods. Bank and Samanta [13], Chakravarthy and Kulshrestha [35], Zhao [161], and Prabhu and Hlynka [111].

1.2.7.4 Transform methods

Transform methods are a group of approaches used in queueing theory to analyze complex systems by converting them into more manageable forms. These techniques reduce the complexity of queueing systems, allowing for easier analysis and solution derivation.

Common transform methods in queueing theory include Laplace transform, Fourier transform, Z–transform, Mellin transform, and the Wiener–Hopf technique. Each method has its own strengths and is suited to different types of queueing problems.

These methods are versatile tools for computing various performance metrics in queueing systems, such as blocking probability, average waiting time, and average queue length. For instance, the Laplace transform is used for finding waiting time distributions in M/M/1 queues, while the Fourier transform is applied to busy period distributions in $M/G/1$ queues.

Recent research has expanded the application of transform methods. Sahakyan et al. [116], Kim et al. [73], Ziolkowski and Tikhonenko [140], Tikhonenko et al. [141], Barbhuiya et al. [14], Tang [137], and Ziolkowski and Tikhonenko [140].

1.2.7.5 Probability generating functions (PGFs)

The probability generating function (PGF) of a discrete random variable X is a power series representation of its probability mass function (PMF). It is defined as:

$$G_{X_t} = E(t^X) = \sum_{x=0}^{\infty} \mathbb{P}(X = x)t^x,$$

where $E[.]$ denotes the expected value operator, $\mathbb{P}(X = x)$ is the PMF of X, and t is a real number.

PGFs are valuable tools for analyzing discrete random variables. The kth derivative of the PGF evaluated at $t = 0$ yields the probability that X takes the value k:

$$\mathbb{P}(X = k) = G_X^{(k)}(0)/k!.$$

Applications of PGFs include finding PMFs and moments of discrete random variables, calculating sums of probabilities, solving recurrence relations, and generating random variables.

Karl Pearson introduced PGFs in the early twentieth century as a method to summarize discrete random variable distributions. Since then, PGFs have become standard tools in probability theory, particularly for discrete distributions and sums of independent random variables.

Recent research has expanded PGF applications in queueing theory. Bouchentouf et al. [28], Bouchentouf et al. [24], Ammar et al. [6], and Bouchentouf et al. [25].

Chapter 2

A Markovian batch arrival queueing system with disasters, working breakdowns, and impatience: mathematical modeling and economic analysis

2.1 Introduction

Over the past two decades, queueing systems with disasters have garnered significant attention due to the rapid development of communication systems and networks. Disasters in these systems lead to the forced departure of all present customers, including the one being served. Such queueing models find applications in various domains. For instance, in computer networks, a virus can act as a delete operation, wiping out all stored data. Extensive research works have been conducted on the subject. Notable contributions include a $M/M/1$ queue with catastrophes by [77], a multi-server retrial queue with negative customers and disasters by [122], and a finite-source discrete-time $Geo/Geo/1$ queue with disasters by [71]. Subsequently, [108] extended the work [71] to $GI/Geo/1$ queues, while [74] investigated $M/G/1$ queues with disasters and working breakdowns. [70] studied $M/G/1$ queues with distastes in a multi-phase random environment and [69] explored $GI/M/1$ queues in multi-phase random environments with disasters and working breakdowns. Recently, [130] discussed a discrete-time $Geo/Geo/1$ queue with feedback, repair and disaster.

Customer behavior, such as balking and reneging, plays a crucial role in real-world queueing systems, where arrivals may be discouraged by long queues. Abundant literature exists on this topic, [156] studied a queue with disasters and impatient customers in which during the breakdown period, the new arrivals become impatient. Then, [30] presented optimal and equilibrium balking strategies in the single server Markovian queue with catastrophes and derived the corresponding Nash equilibrium and social optimal strategies. Later, they analyzed the effect of catastrophes on the strategic customer behavior in queueing systems [31]. Recently, [6] presented the transient analysis of impatient customers in a Markovian single server queue with disasters queue in random environment. Other contributions on impatience customers' in different queueing models can be found in [126, 45, 162, 158, 19, 22, 21, 40].

Queueing systems with batch arrivals have a long history, dating back to the works of [115, 51, 120, 33, 88, 87, 93]. Subsequent researches have explored priority queues [136], server vacations [85], and heavy traffic limit theories [41, 107, 110]. Recent studies have investigated batch arrivals in conjunction with multiple working vacation [11], disasters and vacation [98], retrial queues [143], fluid queues [56], breakdowns and vacation [8, 10], vacation/working vacation queues with impatience [24, 26, 22], and group clearance [36].

This chapter presents a novel contribution by considering an infinite-capacity multi-server Markovian queue with batch arrivals, Bernoulli feedback, working breakdowns, balking, and reneging. This model assumes that service continues at a reduced rate during repair periods, reflecting practical scenarios such as computer networks under virus attacks or machine replacements in manufacturing systems. The main contributions of this work can be summarized as follows:

- Establishing the stability condition for the proposed queueing system.
- Obtaining the steady-state solution for the system by using probability generating functions (PGFs), which provide a powerful approach for analyzing discrete probability distributions and stochastic processes.
- Deriving important performance measures from the steady-state probability distributions.
- Formulating a cost model for the queueing system to conduct an economic analysis.
- Performing a numerical analysis to validate the analytical results and investigate the impact of different system parameters on the performance measures, total expected cost, and total expected profit.

The body of the remainder of this chapter is organized: Description of the system and a practical application of the suggested queueing model are given in Section 2. In Section 3, the analysis of the system is established. In Section 4, we formulated the performance measures of the system. In Section 5, we present a cost model. Then, in Section 6, numerical simulation results are provided and finally, in Section 7, we conclude the work done.

2.2 Model description

An infinite-capacity multi-server queue batch arrivals, Bernoulli feedback, working breakdowns, balking, and reneging is considered:

- Customers arrive in batches according to a Poisson process with rate λ. We consider our system in which the size of an arriving batch is drawn from an independent and identically distributed sequence of random variables. We assume that the times of arrivals are given by a Poisson process. The arrival batch size X is a random variable with probability mass function $\mathbb{P}(X = l) = b_l; l = 1, 2, ...$ They are served in accordance with First Come First Served 'FCFS' discipline.

- The service time during normal busy period are supposed to exponentially distributed with rate μ.

- During the busy period, the system may break down. At this time, all customers present are removed out and the system (all the servers as one station) is sent for a reparation. The inter-arrival times between successive breakdowns are assumed to be distributed exponentially with rate η. Repair times have an exponential distribution with rate ϑ.

- On arrival, if a batch of customers find the c servers busy, they may decide to enter the system with a certain probability θ, or balk wit a complementary probability $\theta' = 1 - \theta$. More precisely, we suppose that the number of customers in the batch is $n(\geq 1)$ then all customers of arrival batch join the system with probability θ, if $n < c$ and all leave the system without receiving service (balk) with probability θ', otherwise.

- During the repair period of the primary servers, new customers can be served by substitute servers. The service times during this period are assumed to be exponentially distributed with a rate ν, where $\nu < \mu$. Once the repair of the servers is completed, the service by the substitute servers is immediately stopped, and the primary servers restart operations at their regular service rate. Additionally, once the system is repaired and the queue becomes empty, all the primary servers return simultaneously to the system, remain idle, and wait for new arrivals.

- During a repair period, customers can get impatient; each customer activates an impatient timer 'T', exponentially distributed with rate χ.

If the customer has not been served before its impatience time has expired he leaves the system without getting a service.

- If a customer is not happy with current service, he can retry many times as a feedback customer with some probability β' or leave the system with a complementary probability β.

- The inter-arrival times, repair times, impatience times, service times are supposed to be mutually independent.

2.2.1 Practical application of the proposed model

The proposed queueing model with batch arrivals, Bernoulli feedback, disasters, working breakdowns, balking, and reneging has practical applications in various manufacturing and production systems, particularly in the electronics industry. Consider a manufacturing facility that produces electronic devices such as smartphones, tablets, or laptops. The devices arrive in batches of random sizes according to a Poisson arrival process and join the queue/server for processing.

The manufacturing system comprises multiple servers, which are specialized machines or workstations responsible for quality checks, testing, and assembly operations on the devices. These servers operate in parallel, and the devices are served following the First-Come First-Served (FCFS) discipline.

However, the system is susceptible to catastrophic events like power failures, fires, or shortage of supplies, which force the machines (servers) to stop their service and evacuate the devices. In such cases, all existing devices in the system are rejected and lost, and the system undergoes a repair process of random duration.

During the repair period, the system can utilize backup generators and emergency staff to provide a substitute service to the arriving devices. However, the service rate of this substitute service is typically lower than the regular service rate. Devices arriving during a normal or breakdown period can decide whether to enter the system or balk (leave without receiving service) based on a certain probability.

Furthermore, devices already in the system during the repair process can also decide whether to stay or leave based on their impatience time. Each device activates an impatience timer with an exponentially distributed duration. If the device's service is not completed before its impatience time expires, it leaves the system without receiving service. Such devices are considered defective products and sent back to the factory.

After receiving service, if a device is not satisfied with the quality or requires additional processing, it can rejoin the queue for another service attempt with a certain feedback probability. This feedback mechanism allows devices to retry the service until they meet the desired quality standards.

2.3 Analysis of the system

2.3.1 Steady-state equations

Let $N(t)$ be the number of customers in the system and let $J(t)$ denote the status of the server at time t. If $J(t) = 1$, the system is functioning, serving customers, whereas if $J(t) = 2$, the system is down, undergoing a repair process.

Let $\{(N(t), J(t)); t \geq 0\}$ represent two-dimensional infinite state continuous-time Markov chain with state space $\mathcal{S} = \{(n, j) : n \geq 0, j = 1, 2\}$.

Let $\pi_{n,j} = \lim_{t\to\infty} \mathbb{P}\{N(t) = n, J(t) = j\}$, $n \geq 0, j = 1, 2$ define the system state probabilities of the process $\{(N(t), J(t)), t \geq 0\}$. Figure 2.1 depicts the state transition diagram of the queueing model under consideration. Then, based on the theory of Markov process, it is easy to show that the steady-state equations of the model are:

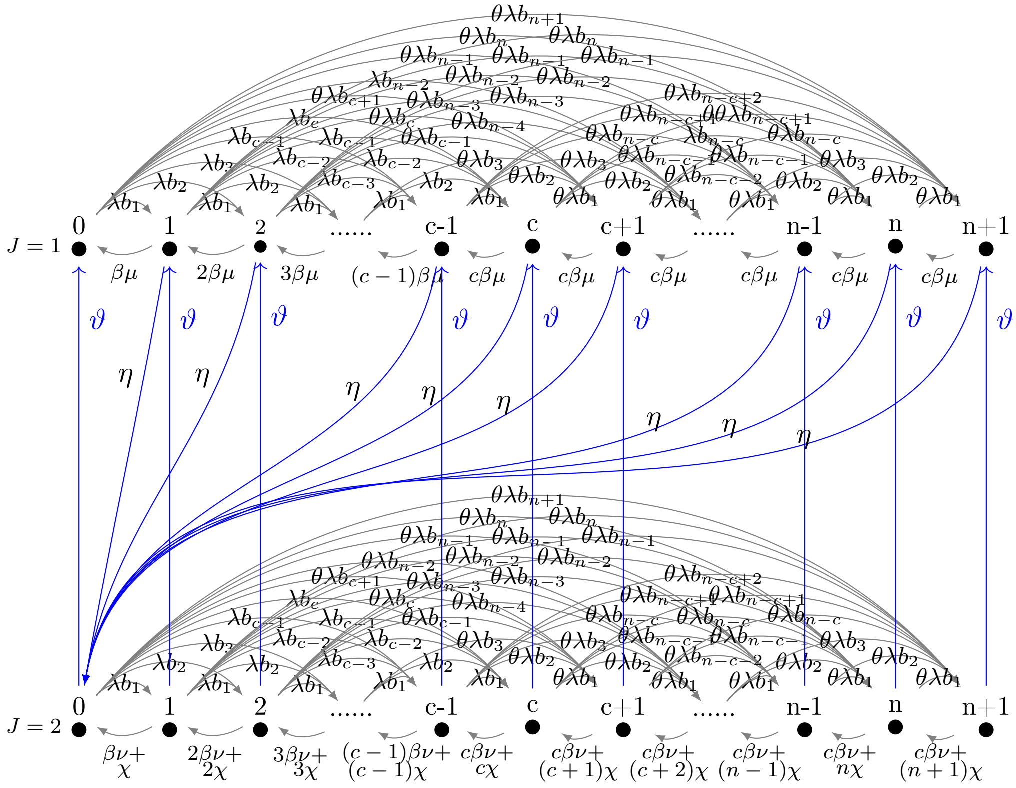

Figure 2.1: State-transition-rate diagram.

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

$$
\begin{aligned}
\lambda\pi_{0,1} &= \vartheta\pi_{0,2} + \beta\mu\pi_{1,1}, \quad n = 0, && (2.1)\\
(\lambda + \eta + \beta\mu)\pi_{1,1} &= \lambda b_1\pi_{0,1} + \vartheta\pi_{1,2} + 2\beta\mu\pi_{2,1}, \quad n = 1, && (2.2)\\
(\lambda + \eta + n\beta\mu)\pi_{n,1} &= \lambda\sum_{m=1}^{n} b_m\pi_{n-m,1} + \vartheta\pi_{n,2} + (n+1)\beta\mu\pi_{n+1,1}, \\
&\quad 2 \le n \le c-1, && (2.3)\\
(\theta\lambda + \eta + n\beta\mu)\pi_{n,1} &= \lambda\sum_{m=1}^{n} b_m\pi_{n-m,1} + \vartheta\pi_{n,2} + c\beta\mu\pi_{n+1,1}, \quad n = c, && (2.4)\\
(\theta\lambda + \eta + c\beta\mu)\pi_{n,1} &= \theta\lambda\sum_{m=1}^{n} b_m\pi_{n-m,1} + \vartheta\pi_{n,2} + c\beta\mu\pi_{n+1,1}, \quad n \ge c, && (2.5)
\end{aligned}
$$

2. If $J(t) = 2$, working breakdown period:

$$
\begin{aligned}
(\lambda + \vartheta)\pi_{0,2} &= \eta\sum_{n=1}^{\infty}\pi_{n,1} + (\beta\nu + \chi)\pi_{1,2}, \quad n = 0, && (2.6)\\
(\lambda + \vartheta + \beta\nu + \chi)\pi_{1,2} &= \lambda b_1\pi_{0,2} + 2(\beta\nu + \chi)\pi_{n+1,2}, \quad n = 1, && (2.7)\\
(\lambda + \vartheta + n(\beta\nu + \chi))\pi_{n,2} &= \lambda\sum_{m=1}^{n} b_m\pi_{n-m,2} + (n+1)(\beta\nu + \chi)\pi_{n+1,2}, \\
&\quad 2 \le n \le c-1, && (2.8)\\
(\theta\lambda + \vartheta + n(\beta\nu + \chi))\pi_{n,2} &= \lambda\sum_{m=1}^{n} b_m\pi_{n-m,2} + (c\beta\nu + (n+1)\chi)\pi_{n+1,2}, \\
&\quad n = c, && (2.9)\\
(\theta\lambda + \vartheta + c\beta\nu + n\chi)\pi_{n,2} &= \theta\lambda\sum_{m=1}^{n} b_m\pi_{n-m,2} + (c\beta\nu + (n+1)\chi)\pi_{n+1,2}, \\
&\quad n \ge c, && (2.10)
\end{aligned}
$$

2.3.2 Stability condition

According to Neuts [103], the infinitesimal generator **Q** for the bivariate process $\{(N(t), J(t)); t \ge 0\}$ is given as follows:

$$
\mathbf{Q}=\left(\begin{array}{ccccccccccc}
\mathbf{A}_0 & \mathbf{C}_1^{(0)} & \mathbf{C}_2^{(0)} & \dots & \dots & \mathbf{C}_c^{(0)} & \mathbf{C}_{c+1}^{(1)} & \dots & \dots & & \\
\mathbf{B}_1 & \mathbf{A}_1 & \mathbf{C}_1^{(0)} & \mathbf{C}_2^{(0)} & \dots & \mathbf{C}_{c-1}^{(0)} & \mathbf{C}_c^{(1)} & \dots & \dots & & \\
\mathbf{D} & \mathbf{B}_2 & \mathbf{A}_2 & \mathbf{C}_1^{(0)} & \dots & \mathbf{C}_{c-2}^{(0)} & \mathbf{C}_{c-1}^{(1)} & \dots & \dots & & \\
\vdots & & \ddots & \ddots & \ddots & & & & & & \\
\mathbf{D} & & & \mathbf{B}_{c-1} & \mathbf{A}_{c-1} & \mathbf{C}_1^{(0)} & \mathbf{C}_2^{(1)} & \dots & \dots & & \\
\mathbf{D} & & & & \mathbf{B}_c & \mathbf{A}_c & \mathbf{C}_1^{(1)} & \dots & \dots & & \\
\mathbf{D} & & & & & \mathbf{B}_{c+1} & \mathbf{A}_{c+1} & \mathbf{C}_1^{(1)} & \dots & & \\
\vdots & & & & & & \ddots & \ddots & \ddots & & \\
\mathbf{D} & & & & & & & \mathbf{B}_N & \mathbf{A}_N & \mathbf{C}_1^{(1)} & \\
\mathbf{D} & & & & & & & & \mathbf{B}_N & \mathbf{A}_N & \mathbf{C}_1^{(1)} \\
\vdots & & & & & & & & & \ddots & \ddots
\end{array}\right),
$$

where N is a sufficiently large number such that when the number of customers $n \geq N$, we approximate the matrices $\mathbf{A}_n$ and $\mathbf{B}_n$ by $\mathbf{A}_N$ and $\mathbf{B}_N$, respectively. In the proposed queueing model, the approximation of the matrices $\mathbf{A}_n$ and $\mathbf{B}_n$ by $\mathbf{A}_N$ and $\mathbf{B}_N$ for $n \geq N$ is employed to facilitate numerical analysis and computation. This approximation is based on the assumption that when the queue length exceeds a certain threshold N, the transition rates within the same level and to the next level can be considered constant, as the dynamics of the system do not significantly change for large queue lengths. The value of N is judiciously chosen such that the difference between the exact and approximated transition rates becomes negligible for $n \geq N$. Specifically, a sufficiently large N is selected, and the matrices $\mathbf{A}_N$ and $\mathbf{B}_N$ are calculated using the steady-state equations and transition rate expressions. Then, for all $n \geq N$, the matrices $\mathbf{A}_n$ and $\mathbf{B}_n$ are approximated by the constant matrices $\mathbf{A}_N$ and $\mathbf{B}_N$, respectively. These approximated matrices are subsequently used in the matrix representation of the Markov chain and in the numerical computations. While this approximation introduces some error, it is a practical consideration that enables the analysis of large-scale queueing systems by improving computational tractability.

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

$$
\mathbf{A}_0=\begin{pmatrix} -\lambda & 0 \\ \vartheta & -(\lambda+\vartheta) \end{pmatrix},\quad \mathbf{C}_l^{(0)}=\begin{pmatrix} \lambda b_l & 0 \\ 0 & \lambda b_l \end{pmatrix},\quad 1 \leq l < c
$$

$$
\mathbf{B}_1=\begin{pmatrix} \beta\mu & \eta \\ 0 & \beta\nu+\chi \end{pmatrix},\quad \mathbf{C}_l^{(1)}=\begin{pmatrix} \theta\lambda b_l & 0 \\ 0 & \theta\lambda b_l \end{pmatrix},\quad l \geq c,
$$

$$
\mathbf{B}_n=\begin{pmatrix} n\beta\mu & 0 \\ 0 & n(\beta\nu+\chi) \end{pmatrix},\quad 2 \leqslant n \leqslant c-1
$$

$$
\mathbf{B}_n=\begin{pmatrix} c\beta\mu & 0 \\ 0 & c\beta\nu+n\chi \end{pmatrix},\quad c \leqslant n \leqslant N-1
$$

$$
\mathbf{B}_n=\begin{pmatrix} c\beta\mu & 0 \\ 0 & c\beta\nu+N\chi \end{pmatrix},\quad n \geq N,\quad \mathbf{D}=\begin{pmatrix} 0 & \eta \\ 0 & 0 \end{pmatrix}
$$

$$\mathbf{A}_n = \begin{pmatrix} -(\lambda + n\beta\mu + \eta) & 0 \\ \vartheta & -(\lambda + n(\beta\nu + \chi) + \vartheta) \end{pmatrix}, \quad 1 \leqslant n \leqslant c-1$$

$$\mathbf{A}_n = \begin{pmatrix} -(\theta\lambda + c\beta\mu + \eta) & 0 \\ \vartheta & -(\theta\lambda + c\beta\nu + n\chi + \vartheta) \end{pmatrix}, \quad c \leqslant n \leqslant N-1$$

$$\mathbf{A}_n = \begin{pmatrix} -(\theta\lambda + c\beta\mu + \eta) & 0 \\ \vartheta & -(\theta\lambda + c\beta\nu + N\chi + \vartheta) \end{pmatrix}, \quad n \geq N$$

In the following Theorem, we present the stability condition of our queueing system.

Theorem 2.1. *The Markov process* $\{(N(t); J(t)), t \geq 0\}$ *is ergodic if and only if*

$$\theta b\lambda < [c\beta\mu\vartheta + (c\beta\nu + N\chi)\eta]\frac{1}{\vartheta + \eta}, \quad \textit{where } b = \sum_{l=1}^{\infty} lb_l. \tag{2.11}$$

Proof. Based on [103], the approximated system is stable and the steady-state probability vector exists if and only if

$$\mathbf{x}\sum_{l=1}^{\infty} l\mathbf{C}_l^{(1)}\mathbf{e}_n < \mathbf{x}\mathbf{B}_N\mathbf{e}_n, \tag{2.12}$$

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

$$\mathbf{F} = \mathbf{D} + \mathbf{B}_N + \mathbf{A}_N + \sum_{l=1}^{\infty} \mathbf{C}_l^{(1)},$$

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

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

Solving the above two equations, we get

$$\mathbf{x} = [x_1, x_2] = \left[\frac{\vartheta}{\vartheta + \eta}, \frac{\eta}{\vartheta + \eta}\right].$$

Then, by substituting $\mathbf{x}$, $\mathbf{e}_n$, $\mathbf{C}_l^{(1)}$, and $\mathbf{B}_N$ into Equation (2.12), we find the stability condition (2.11). □

2.3.3 Analysis of the steady-state probability distribution

Define the probability generating functions as:

$$G_j(z) = \sum_{n=0}^{\infty} \pi_{n,j} z^n, |z| \leq 1, \ j = 1, 2, \ G_j^{'}(z) = \frac{d}{dz} G_j(z) = \sum_{n=1}^{\infty} n\pi_{n,j} z^{n-1}, \ j = 1, 2,$$

and $B(z) = \sum_{n=1}^{\infty} b_n z^n$, with $B(1) = \sum_{n=1}^{\infty} b_n = 1$.
Multiplying Eqs. (2.1)-(2.5) by z^n and summing all possible values of n, we get:

$$\begin{aligned} &[\theta\lambda z(B(z)-1) + c\beta\mu(1-z) - z\eta]\, G_1(z) + z\vartheta G_2(z) = \lambda\theta^{'} z\psi_1(z) \\ &+\beta\mu(1-z)\psi_2(z) - \lambda\theta^{'} z\psi_3(z) - z\eta\pi_{0,1}, \end{aligned} \tag{2.13}$$

where $\psi_1(z) = \sum_{n=0}^{c-1} \pi_{n,1} z^n$, $\psi_2(z) = \sum_{n=0}^{c-1} (c-n)\pi_{n,1} z^n$, and $\psi_3(z) = \sum_{n=1}^{c} \sum_{m=1}^{n} b_m \pi_{n-m,1} z^n$.
Similarly, multiplying Eqs. (2.6)-(2.10) by z^n then summing all possible values of n, we obtain:

$$\begin{aligned} &\chi z(1-z)G_2^{'}(z) + [\theta\lambda z(B(z)-1) + c\beta\nu(1-z) - z\vartheta]G_2(z) = \lambda\theta^{'} z\varphi_1(z) \\ &+\beta\nu(1-z)\varphi_2(z) - \lambda\theta^{'} z\varphi_3(z) - \eta z G_1(1) + z\eta\pi_{0,1}, \end{aligned} \tag{2.14}$$

with $\varphi_1(z) = \sum_{n=0}^{c-1} \pi_{n,2} z^n$, $\varphi_2(z) = \sum_{n=0}^{c-1} (c-n)\pi_{n,2} z^n$, and $\varphi_3(z) = \sum_{n=1}^{c} \sum_{m=1}^{n} b_m \pi_{n-m,2} z^n$.
Next, using the recursive method, we get:

$$\begin{cases} \pi_{n,1} = L_n\pi_{0,1} + R_n\pi_{0,2} + N_n G_1(1), \\ \pi_{n,2} = A_n\pi_{0,2} + B_n\pi_{0,1} + C_n G_1(1), \end{cases}$$

where

$$A_n = \begin{cases} 1, & \text{n=0;} \\ \dfrac{\lambda+\vartheta}{\beta\nu+\chi}, & \text{n=1;} \\ \varpi_n A_{n-1} + \frac{\kappa}{n} \sum_{m=1}^{n-1} b_m A_{n-m-1}, & n \geqslant 2, \end{cases}$$

$$B_n = \begin{cases} 0, & \text{n=0;} \\ \dfrac{\eta}{\beta\nu+\chi}, & \text{n=1;} \\ \varpi_n B_{n-1} + \frac{\kappa}{n} \sum_{m=1}^{n-1} b_m B_{n-m-1}, & n \geqslant 2, \end{cases}$$

$$C_n = \begin{cases} 0, & \text{n=0;} \\ \dfrac{-\eta}{\beta\nu+\chi}, & \text{n=1;} \\ \varpi_n C_{n-1} + \frac{\kappa}{n}\displaystyle\sum_{m=1}^{n-1} b_m C_{n-m-1}, & n \geqslant 2, \end{cases}$$

$$L_n = \begin{cases} 1, & \text{n=0;} \\ \dfrac{\lambda}{\beta\mu}, & \text{n=1;} \\ \zeta_n L_{n-1} + \frac{\kappa^{'}}{n}\displaystyle\sum_{m=1}^{n-1} b_m L_{n-m-1} + \dfrac{R_1}{n} B_{n-1}, & n \geqslant 2, \end{cases}$$

$$R_n = \begin{cases} 0, & \text{n=0;} \\ \dfrac{-\vartheta}{\beta\mu}, & \text{n=1;} \\ \zeta_n R_{n-1} + \frac{\kappa^{'}}{n}\displaystyle\sum_{m=1}^{n-1} b_m R_{n-m-1} + \dfrac{R_1}{n} A_{n-1}, & n \geqslant 2, \end{cases}$$

$$N_n = \begin{cases} 0, & \text{n=0;} \\ 0, & \text{n=1;} \\ \zeta_n N_{n-1} + \frac{\kappa^{'}}{n}\displaystyle\sum_{m=1}^{n-1} b_m N_{n-m-1} + \dfrac{R_1}{n} C_{n-1}, & n \geqslant 2, \end{cases}$$

such that $\kappa = \dfrac{-\lambda}{\beta\nu+\chi}$, $\varpi_n = \dfrac{\lambda+\vartheta}{n(\beta\nu+\chi)} + \dfrac{n-1}{n}$, $\kappa^{'} = \dfrac{-\lambda}{\beta\mu}$, and $\zeta_n = \dfrac{\lambda+\eta}{n\beta\mu} + \dfrac{n-1}{n}$.

Further, for $z \neq 1$ *and* $z \neq 0$, Eqs. (2.13) and (2.14) can be respectively written as:

$$\begin{aligned} G_1(z) = -\frac{z\vartheta}{\xi(z)} G_2(z) &+ \left[\frac{\lambda\theta^{'} z(L_1(z) - L_0(z)) + \beta\mu(1-z)L_2(z) - z\eta}{\xi(z)}\right]\pi_{0,1} \\ &+ \left[\frac{\lambda\theta^{'} z(R_1(z) - R_0(z)) + \beta\mu(1-z)R_2(z)}{\xi(z)}\right]\pi_{0,2} \\ &+ \left[\frac{\lambda\theta^{'} z(N_1(z) - N_0(z)) + \beta\mu(1-z)N_2(z)}{\xi(z)}\right] G_1(1), \end{aligned} \tag{2.15}$$

where $\xi(z) = \theta\lambda z(B(z)-1) + c\beta\mu(1-z) - z\eta$,

$$L_0(z) = \sum_{n=1}^{c}\sum_{m=1}^{n} b_m L_{n-m} z^n,\ \ L_1(z) = \sum_{n=0}^{c-1} L_n z^n,\ \ L_2(z) = \sum_{n=0}^{c-1}(c-n)L_n z^n,$$
$$R_0(z) = \sum_{n=1}^{c}\sum_{m=1}^{n} b_m R_{n-m} z^n,\ \ R_1(z) = \sum_{n=0}^{c-1} R_n z^n,\ \ R_2(z) = \sum_{n=0}^{c-1}(c-n)R_n z^n,$$
$$N_0(z) = \sum_{n=1}^{c}\sum_{m=1}^{n} b_m N_{n-m} z^n,\ \ N_1(z) = \sum_{n=0}^{c-1} N_n z^n,\ \ N_2(z) = \sum_{n=0}^{c-1}(c-n)N_n z^n,$$

and

$$G_2'(z) + \left[\frac{\theta\lambda}{\chi}H'(z) + \frac{c\beta\nu}{z\chi} - \frac{\vartheta}{\chi(1-z)}\right]G_2(z) = \left[\frac{\theta'\lambda}{\chi(1-z)}\left[A_1(z)-A_0(z)\right] + \frac{\beta\nu}{z\chi}A_2(z)\right]\pi_{0,2}$$
$$+\left[\frac{\theta'\lambda}{\chi(1-z)}\left[B_1(z)-B_0(z)\right] + \frac{\beta\nu}{z\chi}B_2(z) + \frac{\eta}{\chi(1-z)}\right]\pi_{0,1}$$
$$+\left[\frac{\theta'\lambda}{\chi(1-z)}\left[C_1(z)-C_0(z)\right] + \frac{\beta\nu}{z\chi}C_2(z) - \frac{\eta}{\chi(1-z)}\right]G_1(1), \tag{2.16}$$

with

$$A_0(z) = \sum_{n=1}^{c}\sum_{m=1}^{n} b_m A_{n-m} z^n,\ \ A_1(z) = \sum_{n=0}^{c-1} A_n z^n,\ \ A_2(z) = \sum_{n=0}^{c-1}(c-n)A_n z^n,$$
$$B_0(z) = \sum_{n=1}^{c}\sum_{m=1}^{n} b_m B_{n-m} z^n,\ \ B_1(z) = \sum_{n=0}^{c-1} B_n z^n,\ \ B_2(z) = \sum_{n=0}^{c-1}(c-n)B_n z^n,$$
$$C_0(z) = \sum_{n=1}^{c}\sum_{m=1}^{n} b_m C_{n-m} z^n,\ \ C_1(z) = \sum_{n=0}^{c-1} C_n z^n,\ \ C_2(z) = \sum_{n=0}^{c-1}(c-n)C_n z^n,$$

and $H(z) = \int_0^z \frac{B(x)-1}{1-x}dx$, $H'(z) = \frac{B(z)-1}{1-z}$.

To solve Eq. (2.14), we multiply both sides of Eq. (2.16) by $e^{\frac{\lambda\theta}{\chi}H(z)}(1-z)^{\frac{\vartheta}{\chi}}z^{\frac{c\beta\nu}{\chi}}$, then we get:

$$G_2(z) = \frac{e^{-\frac{\lambda\theta}{\chi}H(z)}}{(1-z)^{\frac{\vartheta}{\chi}}z^{\frac{c\beta\nu}{\chi}}}\Bigg\{\left(\frac{\theta'\lambda}{\chi}K_0(z) + \frac{\beta\nu}{\chi}K_1(z)\right)\pi_{0,2}+$$
$$\left(\frac{\theta'\lambda}{\chi}K_2(z) + \frac{\beta\nu}{\chi}K_3(z) + \frac{\eta}{\chi}K_4(z)\right)\pi_{0,1} + \left[\frac{\theta'\lambda}{\chi}K_5(z) + \frac{\beta\nu}{\chi}K_6(z) - \frac{\eta}{\chi}K_4(z)\right]G_1(1)\Bigg\}, \tag{2.17}$$

where

$$K_0(z) = \int_0^z e^{\frac{\lambda\theta}{\chi}H(x)}(1-x)^{\frac{\vartheta}{\chi}-1}x^{\frac{c\beta\nu}{\chi}}\left(A_1(x)-A_0(x)\right)dx,$$

$$K_1(z) = \int_0^z e^{\frac{\lambda\theta}{\chi}H(x)}(1-x)^{\frac{\vartheta}{\chi}}x^{\frac{c\beta\nu}{\chi}-1}A_2(x)dx,$$
$$K_2(z) = \int_0^z e^{\frac{\lambda\theta}{\chi}H(x)}(1-x)^{\frac{\vartheta}{\chi}-1}x^{\frac{c\beta\nu}{\chi}}\left(B_1(x)-B_0(x)\right)dx,$$
$$K_3(z) = \int_0^z e^{\frac{\lambda\theta}{\chi}H(x)}(1-x)^{\frac{\vartheta}{\chi}}x^{\frac{c\beta\nu}{\chi}-1}B_2(x)dx,$$
$$K_4(z) = \int_0^z e^{\frac{\lambda\theta}{\chi}H(x)}(1-x)^{\frac{\vartheta}{\chi}-1}x^{\frac{c\beta\nu}{\chi}}dx,$$
$$K_5(z) = \int_0^z e^{\frac{\lambda\theta}{\chi}H(x)}(1-x)^{\frac{\vartheta}{\chi}-1}x^{\frac{c\beta\nu}{\chi}}\left(C_1(x)-C_0(x)\right)dx,$$
$$K_6(z) = \int_0^z e^{\frac{\lambda\theta}{\chi}H(x)}(1-x)^{\frac{\vartheta}{\chi}}x^{\frac{c\beta\nu}{\chi}-1}C_2(x)dx.$$

Taking limit as $z \to 1$ in Eq. (2.17) we get:

$$G_2(1) = e^{-\frac{\lambda\theta}{\chi}H(1)}[\left(\frac{\theta'\lambda}{\chi}K_0(1)+\frac{\beta\nu}{\chi}K_1(1)\right)\pi_{0,2}+\left(\frac{\theta'\lambda}{\chi}K_2(1)+\frac{\beta\nu}{\chi}K_3(1)+\frac{\eta}{\chi}K_4(1)\right)\pi_{0,1}$$
$$+\left(\frac{\theta'\lambda}{\chi}K_5(1)+\frac{\beta\nu}{\chi}K_6(1)-\frac{\eta}{\chi}K_4(1)\right)G_1(1)]\lim_{z\to 1}(1-z)^{\frac{-\vartheta}{\chi}}z^{\frac{-c\beta\nu}{\chi}}. \tag{2.18}$$

Since $G_2(1) = \sum_{n=0}^{\infty}\pi_{n,2} > 0$ and $\lim_{z\to 1}(1-z)^{\frac{-\vartheta}{\chi}}z^{\frac{-c\beta\nu}{\chi}} = \infty$ we must have that :

$$\left(\frac{\theta'\lambda}{\chi}K_0(1)+\frac{\beta\nu}{\chi}K_1(1)\right)\pi_{0,2}+\left(\frac{\theta'\lambda}{\chi}K_2(1)+\frac{\beta\nu}{\chi}K_3(1)+\frac{\eta}{\chi}K_4(1)\right)\pi_{0,1}$$
$$+\left(\frac{\theta'\lambda}{\chi}K_5(1)+\frac{\beta\nu}{\chi}K_6(1)-\frac{\eta}{\chi}K_4(1)\right)G_1(1) = 0. \tag{2.19}$$

Therefore

$$G_1(1) = \Theta_1\pi_{0,2}+\Theta_2\pi_{0,1}, \tag{2.20}$$

where

$$\Theta_1 = \frac{-\theta'\lambda K_0(1)-\beta\nu K_1(1)}{\theta'\lambda K_5(1)+\beta\nu K_6(1)-\eta K_4(1)},$$

and

$$\Theta_2 = \frac{-\eta K_4(1)-\theta'\lambda K_2(1)-\beta\nu K_3(1)}{\theta'\lambda K_5(1)+\beta\nu K_6(1)-\eta K_4(1)}.$$

Further, by taking $z = 1$ in Eq. (2.15), we find

$$G_2(1) = \left[\frac{\eta + \lambda\theta^{'}(N_1(1) - N_0(1))}{\vartheta}\right] G_1(1) + \left[\frac{\lambda\theta^{'}(L_1(1) - L_0(1)) - \eta}{\vartheta}\right] \pi_{0,1} + \left[\frac{\lambda\theta^{'}(R_1(1) - R_0(1))}{\vartheta}\right] \pi_{0,2}. \tag{2.21}$$

By substituting Eq. (2.20) into Eq. (2.21) we obtain:

$$G_2(1) = \Psi_1 \pi_{0,2} + \Psi_2 \pi_{0,1}, \tag{2.22}$$

where

$$\begin{aligned} \Psi_1 &= \frac{(\eta + \lambda\theta^{'}(N_1(1) - N_0(1)))\Theta_1}{\vartheta} + \frac{\lambda\theta^{'}(R_1(1) - R_0(1))}{\vartheta}, \\ \Psi_2 &= \frac{(\eta + \lambda\theta^{'}(N_1(1) - N_0(1)))\Theta_2}{\vartheta} + \frac{\lambda\theta^{'}(L_1(1) - L_0(1)) - \eta}{\vartheta}. \end{aligned}$$

Next, by taking $z = 1$ in Eqs. (2.13)-(2.14), we respectively have:

$$-\eta G_1(1) + \vartheta G_2(1) = \lambda\theta^{'}(\psi_1(1) - \psi_3(1)) - \eta\pi_{0,1}, \tag{2.23}$$

$$\eta G_1(1) - \vartheta G_2(1) = \lambda\theta^{'}(\varphi_1(1) - \varphi_3(1)) + \eta\pi_{0,1}. \tag{2.24}$$

Summing both (2.23) and (2.24) we obtain:

$$\psi_1(1) - \psi_3(1) = \varphi_3(1) - \varphi_1(1), \tag{2.25}$$

where
$\varphi_3(1) = A_0(1)\pi_{0,2} + B_0(1)\pi_{0,1} + C_0(1)G_1(1)$,
$\varphi_1(1) = A_1(1)\pi_{0,2} + B_1(1)\pi_{0,1} + C_1(1)G_1(1)$,
$\psi_3(1) = L_0(1)\pi_{0,1} + R_0(1)\pi_{0,2} + N_0(1)G_1(1)$,
$\psi_1(1) = L_1(1)\pi_{0,1} + R_1(1)\pi_{0,2} + N_1(1)G_1(1)$.

Further, by substituting Eq. (2.25) into (2.20) we find:

$$\pi_{0,2} = \Gamma(1)\pi_{0,1}, \tag{2.26}$$

where,

$$\Gamma(1) = \frac{L_0(1) - L_1(1) + B_0(1) - B_1(1) - \Theta_2\left[C_1(1) - C_0(1) + N_1(1) - N_0(1)\right]}{A_1(1) - A_0(1) + R_1(1) - R_0(1) + \Theta_1\left[C_1(1) - C_0(1) + N_1(1) - N_0(1)\right]}.$$

The following Theorem presents the steady-state probabilities of the considered queueing system.

Theorem 2.2. *Under the stability condition, the steady-state probabilities are given by*

$$\pi_{.,1} = G_1(1) = [\Theta_1\Gamma(1) + \Theta_2]\,\pi_{0,1}, \tag{2.27}$$

and

$$\pi_{.,2} = G_2(1) = [\Psi_1\Gamma(1) + \Psi_2]\,\pi_{0,1}, \tag{2.28}$$

where

$$\pi_{0,1} = \left\{\Theta_1\Gamma(1) + \Theta_2 + \Psi_1\Gamma(1) + \Psi_2\right\}^{-1}.$$

Proof. By substituting Eq. (2.26) into Eqs. (2.20) and (2.22) we get $\pi_{.,1}$ and $\pi_{.,2}$, respectively.
Then, using the normalization condition:

$$\sum_{n=0}^{\infty}\sum_{j=1}^{2}\pi_{n,j} = 1 \Leftrightarrow \pi_{.,1} + \pi_{.,2} = 1, \text{ we obtain } \pi_{0,1}.$$

□

Let L denote the number of customers in the system. Then we have $E(L) = E(L_1) + E(L_2)$, where $E(L_1)$ is the mean system size when the system is on busy period and $E(L_2)$ represents the mean system size when the system is on working repair.

Theorem 2.3. *The mean system sizes during busy and working breakdown periods can be expressed as*

$$E(L_1) = \frac{1}{\eta}\left[(\theta\lambda B^{'}(1) - c\beta\mu - \eta)(\Theta_1\Gamma(1) + \Theta_2) + \vartheta(\Psi_1\Gamma(1) + \Psi_2) + \eta\right]\pi_{0,1}$$
$$+\frac{1}{\eta}\left[\lambda\theta^{'}(\psi_3(1) + \psi_3^{'}(1) - \psi_1(1) - \psi_1^{'}(1)) + \beta\mu\psi_2(1) + \vartheta E(L_2)\right],$$

and

$$E(L_2) = \frac{1}{\chi + \vartheta}[\left[\left(\theta\lambda B^{'}(1) - c\beta\nu - \vartheta\right)[\Psi_1\Gamma(1) + \Psi_2] - \eta + \eta(\Theta_1\Gamma(1) + \Theta_2)\right]\pi_{0,1}$$
$$+\lambda\theta^{'}(\varphi_3(1) + \varphi_3^{'}(1) - \varphi_1(1) - \varphi_1^{'}(1)) + \beta\nu\varphi_2(1)].$$

Proof. We derive the equation Eq. (2.13) then setting $z \longrightarrow 1$ and , we get:

$$\left[\theta\lambda B^{'}(1) - c\beta\mu - \eta\right]G_1(1) + \vartheta G_2(1) + \vartheta G_2^{'}(1) - \eta G_1^{'}(1)$$
$$= \lambda\theta^{'}\left[-\psi_3(1) - \psi_3^{'}(1) + \psi_1(1) + \psi_1^{'}(1)\right] - \beta\mu\psi_2(1) - \eta\pi_{0,1},$$

where $G_1(1)$ and $G_2(1)$ are given in Eqs. (2.27) and (2.28), respectively. Therefore

$$E(L_1) = \lim_{z\to 1} G_1^{'}(z)$$
$$= \frac{1}{\eta}\left[(\theta\lambda B^{'}(1) - c\beta\mu - \eta)(\Theta_1\Gamma(1) + \Theta_2) + \vartheta(\Psi_1\Gamma(1) + \Psi_2) + \eta\right]\pi_{0,1}$$
$$+\frac{1}{\eta}\left[\lambda\theta^{'}(\psi_3(1) + \psi_3^{'}(1) - \psi_1(1) - \psi_1^{'}(1)) + \beta\mu\psi_2(1) + \vartheta E(L_2)\right].$$

Next, differentiating Eq. (2.14) and taking $z \longrightarrow 1$ we find:

$$(\chi+\vartheta)G_2^{'}(1) = \left[\theta\lambda B^{'}(1) - c\beta\nu - \vartheta\right] G_2(1) + \lambda\theta^{'}(\varphi_3(1) + \varphi_3^{'}(1) - \varphi_1(1) - \varphi_1^{'}(1)) + \beta\nu\varphi_2(1) - \eta\left[\pi_{0,1} - G_1(1)\right].$$

Thus

$$\begin{aligned} E(L_2) &= \lim_{z\to 1} G_2^{'}(z) \\ &= \frac{1}{\chi+\vartheta}[\left[\left(\theta\lambda B^{'}(1) - c\beta\nu - \vartheta\right)[\Psi_1\Gamma(1) + \Psi_2] - \eta + \eta(\Theta_1\Gamma(1) + \Theta_2)\right]\pi_{0,1} \\ &\quad + \lambda\theta^{'}(\varphi_3(1) + \varphi_3^{'}(1) - \varphi_1(1) - \varphi_1^{'}(1)) + \beta\nu\varphi_2(1)]. \end{aligned}$$

□

2.4 Performance measures and cost model

2.4.1 Performance measures

In this subpart of paper, useful performance measures are presented.

Corollary 2.4. *The mean number of customers in the queue is given as:*

$$E(L_q) = \sum_{n=c+1}^{\infty} (n-c)\left(\pi_{n,1} + \pi_{n,2}\right) = E(L) - c + \psi_2(1) + \varphi_2(1).$$

Corollary 2.5. *1. The probability that the servers are in working repair period is presented as:*

$$P_{wr} = G_2(1) = \sum_{n=0}^{\infty} \pi_{n,2}.$$

2. The probability that the servers are in a normal busy period is presented as:

$$P_b = 1 - P_{wr}.$$

3. The probability that the servers are working either during busy or repair period is presented as:

$$P_w = \sum_{n=1}^{\infty}(\pi_{n,2} + \pi_{n,1}).$$

Corollary 2.6.

1. The mean number of customers served per unit time is given as:

$$\begin{aligned} N_s &= \beta\mu\sum_{n=1}^{c-1} n\pi_{n,1} + c\beta\mu\sum_{n=c}^{\infty}\pi_{n,1} + \beta\nu\sum_{n=1}^{c-1} n\pi_{n,2} + c\beta\nu\sum_{n=c}^{\infty}\pi_{n,2} \\ &= \beta\mu\left[cP_{busy} - \psi_2(1)\right] + \beta\nu\left[cP_{wr} - \varphi_2(1)\right]. \end{aligned}$$

2. The average rate of abandonment of customers due to impatience is given as:

$$R_a = \chi E(L_2).$$

3. The average rate of balking is given as:

$$R_{balk} = \lambda(1-\theta)\left[\sum_{n=c}^{\infty}\pi_{n,1} + \sum_{n=c}^{\infty}\pi_{n,2}\right] = \lambda(1-\theta)\left[1-\psi_1(1)-\varphi_1(1)\right].$$

2.4.2 Cost model

Cost-profit analysis is very beneficial in the application of real-life situations arising from industrial and technical situations.

The total expected cost (T_{cost}) is defined as:

$$\begin{aligned} T_{\text{cost}} &= C_bP_b + C_{rp}P_{wr} + (C_lE(L)) + (C_rR_{ren}) + c(\mu+\nu) \\ &\quad \times(C_s + (1-\beta)C_f) + cC_p, \end{aligned}$$

where C_b is the cost per unit time during normal busy period, C_{rp}; the cost per unit time during working repair period, C_l; the holding cost per unit time, C_r; the cost per unit time when a customer is lost due to impatience, C_s; the cost per service per unit time, C_f; the cost per unit time when a customer returns to the system as a feedback, and C_p; the fixed server purchase cost per unit.

Let $\mathcal{R}$ be the revenue earned for providing service to a customer, then the total expected revenue per unit time (T_{revenue}) of the system is as:

$$T_{\text{revenue}} = R \times N_s.$$

The total expected profit (T_{profit}) per unit time of the system is as:

$$T_{\text{profit}} = T_{\text{revenue}} - T_{\text{cost}}.$$

2.5 Numerical analysis

To validate the analytical results obtained through mathematical modeling and analysis, we employ computational techniques. These methods allow us to compute and approximate the relevant quantities, including steady-state probabilities and performance measures for the manufacturing system. By comparing the numerically obtained results with the analytical expressions, we validate the accuracy of our derived solutions.

While numerical analysis introduces approximations and potential numerical errors, it complements analytical methods by providing a practical means of verifying results and acquiring a deeper understanding of the behavior of complex queueing systems.

In this section, important numerical results are presented in the form of Tables and Graphs in order to illustrate the effect of various system parameters on different system characteristics, (T_{cost}) and (T_{profit}), using R program. The arrival batch size X follows a geometric distribution with parameter σ ,that is $P(X = l) = (1-\sigma)^{l-1}\sigma$, with $0 < \sigma < 1$, and $l = 1, 2, \ldots$. Therefore, $B(z) = \dfrac{\sigma z}{1-(1-\sigma)z}$.

For our analysis, we consider the manufacturing system discussed above. Unless their values are indicated in the appropriate places, the model parameters are assumed to be as follows: the devices arrive in groups of random size according to a Poisson arrival process with rate $\lambda = 1.0$ devices per minute, and join the queue/server for processing. The system has $c = 3$ primary servers. The service time of each machine is exponentially distributed with rate $\mu = 1.9$ devices per minute. The time between successive breakdowns is exponentially distributed with rate $\eta = 5.0$ breakdowns per minute, and the repair time is exponentially distributed with rate $\vartheta = 2.0$ repairs per minute. During a breakdown, the substitute service time is exponentially distributed with rate $\nu = 0.5$ devices per minute, where $\nu < \mu$. The devices can decide whether to enter the system or not, based on the probability $\theta = 0.8$. The devices that are already in the system can also decide whether to stay or leave, based on their impatience time, which is exponentially distributed with rate $\chi = 0.2$ devices per minute. If a device gets a service but is not satisfied, it can retry the service with probability $\beta' = 0.5$ or leave the system with probability $\beta = 0.5$, and $\sigma = 0.7$.

To evaluate the cost and revenue of the system, we can use the following parameters: we take the cost parameters as $C_b = \$3.5$, $C_{rp} = \$2$, $C_l = \$2.5$, $C_r = \$2$, $C_s = \$0.11$, $C_f = \$0.11$, $C_p = \$1$, and $R = \$70$. These values can be adjusted according to the market conditions and the quality of the devices. Numerical results are presented in Table 2.5 and Figs. 2.2-2.7:

		$\pi_{0,1}$	P_{wr}	P_b	$E(L_1)$	$E(L_2)$	$E(L)$	R_{balk}	R_{ren}	N_s
	1.0	0.5156	0.3445	0.6555	0.1448	0.1783	0.3231	0.1238	0.0357	0.4922
λ	1.5	0.3815	0.4387	0.5613	0.2163	0.3288	0.5451	0.2434	0.0658	0.6480
	2.0	0.2894	0.5025	0.4975	0.2843	0.4869	0.7712	0.3942	0.0974	0.7749
	1.1	0.2873	0.5041	0.4959	0.3465	0.4884	0.8349	0.3831	0.0977	0.4797
μ	1.5	0.2889	0.5029	0.4971	0.3160	0.4872	0.8033	0.3886	0.0974	0.6244
	1.9	0.2894	0.5025	0.4975	0.2843	0.4869	0.7712	0.3942	0.0974	0.7749
	0.3	0.2874	0.5040	0.4960	0.2903	0.5051	0.7953	0.3990	0.1010	0.7349
ν	0.4	0.2884	0.5033	0.4967	0.2872	0.4959	0.7831	0.3966	0.0992	0.7553
	0.5	0.2894	0.5025	0.4975	0.2843	0.4869	0.7712	0.3942	0.0974	0.7749
	0.5	0.2894	0.5025	0.4975	0.2843	0.4869	0.7712	0.3942	0.0974	0.7749
θ	0.7	0.2874	0.5060	0.4940	0.3270	0.5334	0.8604	0.2391	0.1067	0.7740
	0.9	0.2855	0.5094	0.4906	0.3705	0.5830	0.9535	0.0805	0.1166	0.7731
	5	0.2894	0.5025	0.4975	0.2843	0.4869	0.7712	0.3942	0.0974	0.7749
η	7	0.3050	0.5364	0.4636	0.2251	0.5197	0.7448	0.3628	0.1039	0.6594
	9	0.3143	0.5575	0.4425	0.1853	0.5402	0.7254	0.3443	0.1080	0.5901
	1	0.1362	0.7108	0.2892	0.2064	1.0894	1.2958	0.5284	0.2179	0.8523
ϑ	1.5	0.2210	0.5926	0.4074	0.2547	0.7019	0.9566	0.4450	0.1404	0.8075
	2	0.2894	0.5025	0.4975	0.2843	0.4869	0.7712	0.3942	0.0974	0.7749
	0.1	0.2874	0.5040	0.4960	0.2914	0.5078	0.7991	0.3990	0.0508	0.7796
χ	0.15	0.2884	0.5033	0.4967	0.2877	0.4971	0.7848	0.3966	0.0746	0.7773
	0.20	0.2894	0.5025	0.4975	0.2843	0.4869	0.7712	0.3942	0.0974	0.7749
	3	0.2894	0.5025	0.4975	0.2843	0.4869	0.7712	0.3942	0.0974	0.7749
c	5	0.3032	0.4959	0.5041	0.3723	0.5333	0.9056	0.1447	0.1067	0.6822
	7	0.3026	0.4975	0.5025	0.4112	0.5616	0.9728	0.0501	0.1123	0.6175
	0.5	0.2894	0.5025	0.4975	0.2843	0.4869	0.7712	0.3942	0.0974	0.7749
β	0.7	0.2898	0.5022	0.4978	0.2154	0.4709	0.6862	0.4006	0.0942	1.1139
	0.9	0.2871	0.5041	0.4959	0.1425	0.4577	0.6002	0.4075	0.0915	1.4710
	0.5	0.2795	0.5030	0.4970	0.3668	0.5924	0.9592	0.5439	0.1185	0.9595
σ	0.7	0.2894	0.5025	0.4975	0.2843	0.4869	0.7712	0.3942	0.0974	0.7749
	0.9	0.3011	0.4981	0.5019	0.2474	0.4180	0.6654	0.2628	0.0836	0.6098

Table 2.1: Effect of λ, μ, ν, ϑ, η, σ, χ, and β on performance measures.

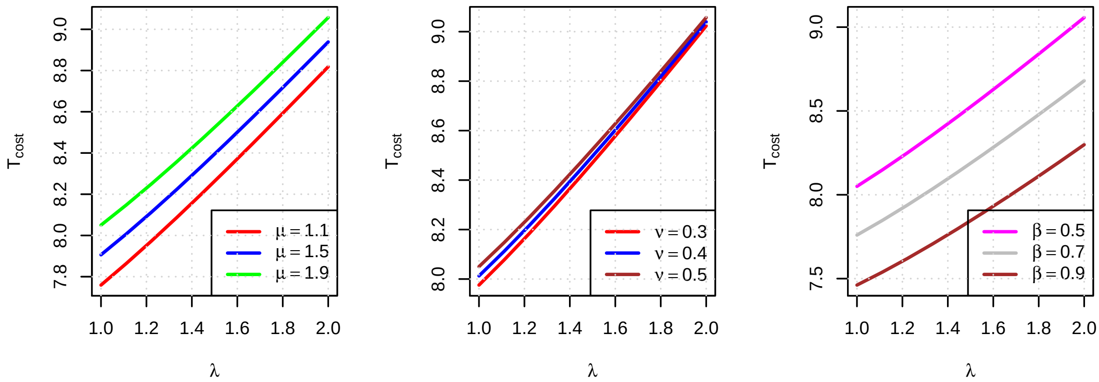

Figure 2.2: Total expected profit versus vs. μ, ν, and β

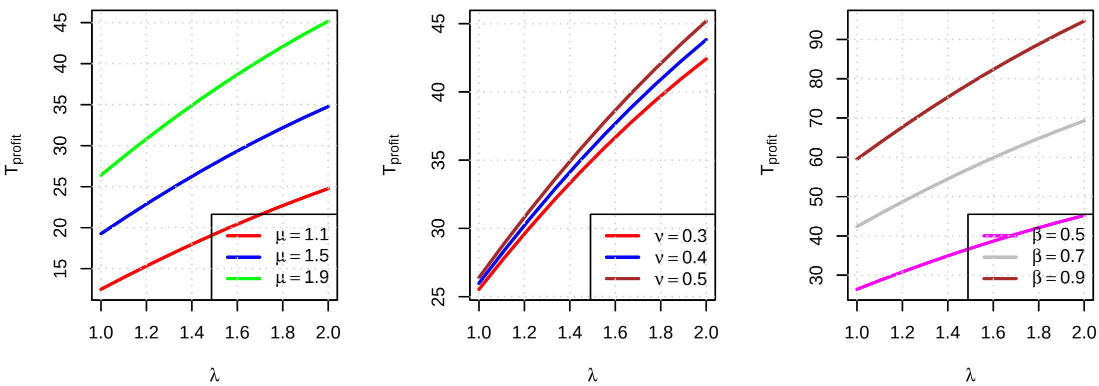

Figure 2.3: Total expected profit versus μ, ν, and β

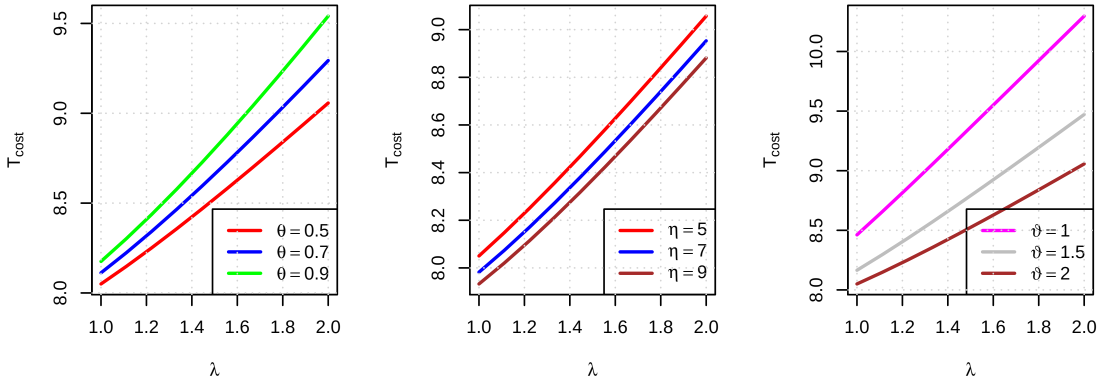

Figure 2.4: Total expected cost versus vs. θ, η, and ϑ

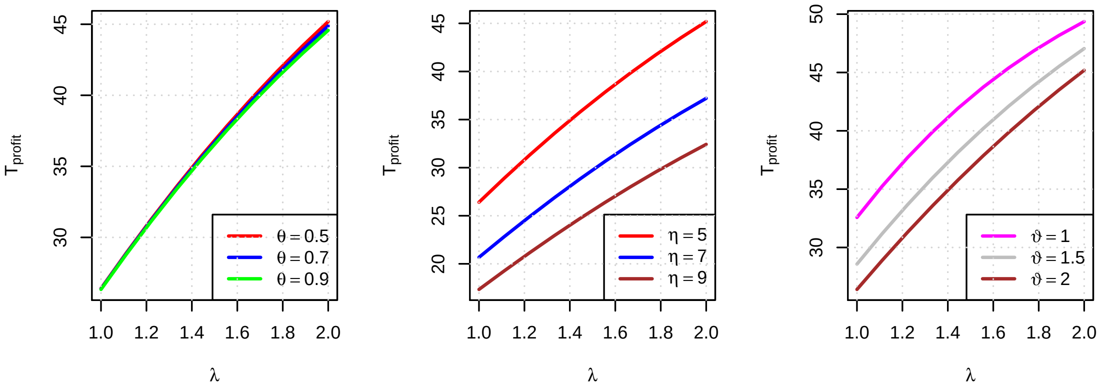

Figure 2.5: Total expected profit versus vs. θ, η, and ϑ

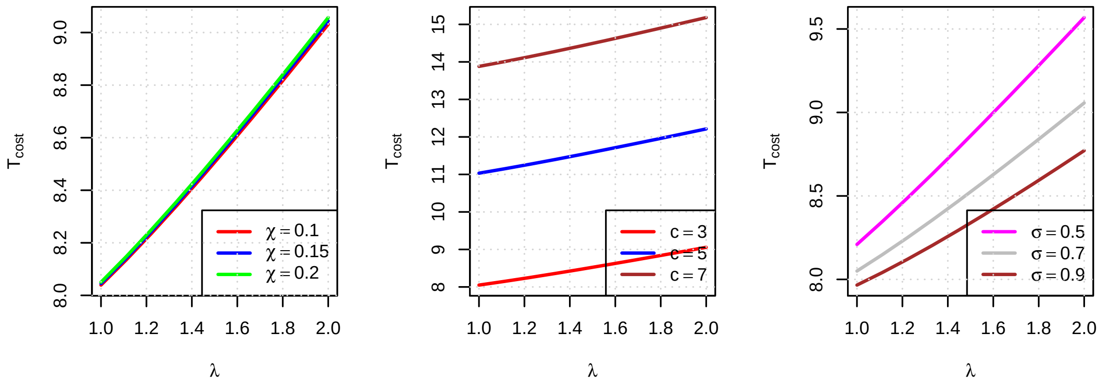

Figure 2.6: Total expected cost versus vs. χ, c, and σ

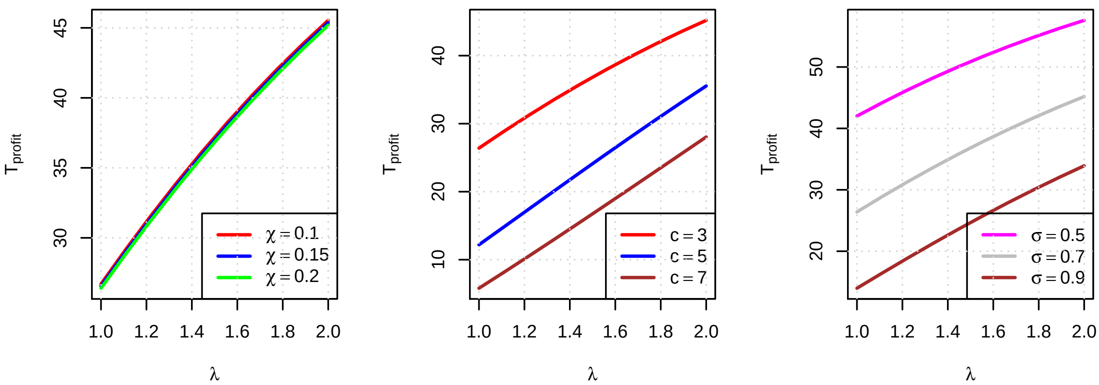

Figure 2.7: Total expected profit versus vs. χ, c, and σ

2.5.1 Results discussion and managerial insights

The numerical experiments explored the sensitivity of various system parameters on performance measures and the cost-profit model. The key observations and potential managerial recommendations are as follows:

1. Arrival rate (λ): A higher arrival rate leads to lower idle probability ($\pi_{0,1}$) and higher expected system lengths ($E(L_1), E(L_2), E(L)$), average balking and reneging rates (R_{balk}, R_{ren}), and mean number of

customers served (N_s). Consequently, both the total expected cost and profit ($T_{\text{cost}}, T_{\text{profit}}$) increase. While higher demand can boost revenue, managers must judiciously balance it against the potential impact on congestion, service quality degradation, and increased system breakdowns.

2. Service rates (μ, ν): Higher service rates during busy and breakdown periods reduce average system lengths and reneging rates, improving customer satisfaction and loyalty. Furthermore, it increases idle probability ($\pi_{0,1}$), mean number of customers served (N_s), and lowers breakdown probability (P_{wr}), enhancing profitability. However, the associated costs and feasibility constraints of increasing service rates should be carefully evaluated.

3. Non-balking probability (θ): A higher non-balking probability attracts more customers, increasing the busy period's average system length ($E(L_1)$) and breakdown probability (P_{wr}). This leads to higher customer losses (R_{ren}), reducing the mean number of customers served and the total expected profit (T_{profit}) while increasing the total expected cost (T_{cost}). Managers must strike a balance between attracting customers and managing congestion, impatience, and system breakdowns.

4. Impatience rate (χ): A higher impatience rate adversely impacts the profit by reducing the mean number of customers served due to reneging. To mitigate this, the manufacturing system should aim to decrease the impatience rate of the devices as much as possible, by providing substitute services, priority queues, or compensation schemes. This can improve customer retention and overall satisfaction.

5. Failure rate (η): A higher failure rate decreases the average system length during the busy period ($E(L_1)$), leading to more customer losses due to impatience (R_{ren}), significantly reducing the total expected profit (T_{profit}). This means that the manufacturing system can reduce the impact of catastrophic events by decreasing the failure rate and increasing the repair rate.

6. Repair rate (ϑ): Even when considering the server's ability to work during breakdowns, a higher repair rate leads to a reduction in the number of customers served and profitability. This can be explained by the fact that the reduced service rate can, sometimes, act as a bottleneck, leading to longer queues and waiting times. Consequently, fewer customers may be served overall, resulting in lower profitability despite the reduction in complete downtime. Therefore, managers must carefully evaluate the cost-benefit trade-off of investing in higher repair rates against the potential revenue gains from reduced complete

downtime and the potential revenue losses due to the server's reduced service capacity during breakdowns.

7. Non-feedback probability (β): Similar to service rates, a higher non-feedback probability leads to an increase in the total expected profit. This is likely due to a reduction in congestion and a more efficient utilization of system resources when fewer customers return to the system. However, a high non-feedback probability may also indicate underlying issues with service quality and customer dissatisfaction. While increasing the feedback probability can boost revenue in the short term by accommodating more retries, it may also exacerbate congestion and waiting times. To strike a balance and maintain long-term profitability, managers should focus on improving service quality through better service rates or providing compensations to enhance overall customer satisfaction. This can help retain customers and mitigate the potential negative impact of a high non-feedback probability on future business.

8. Batch size probability (σ): A higher probability of smaller batch sizes (σ closer to 1) decreases the mean number of customers in the system, customers served, the average rates of balking and reneging, probability of breakdowns, total expected cost, and total expected profit. Attracting more devices by favoring smaller batches can be a viable strategy. However, this approach also increases congestion and customer waiting times. Managers must carefully balance the trade-off between batch size probability and waiting times, as a high probability of smaller batches can exacerbate impatience and dissatisfaction among customers.

9. The number of servers (c): A higher number of servers in the system leads to higher total expected costs and lower total expected profits. This counterintuitive result can be attributed to potential inefficiencies and coordination challenges associated with managing a larger number of servers. With more servers, there is an increased likelihood of underutilization or imbalanced workload distribution, leading to inefficient resource utilization and longer waiting times for customers. Additionally, a higher number of servers may also increase the system's complexity, potentially leading to more frequent breakdowns or maintenance requirements, further contributing to reduced throughput and profitability. Therefore, managers should carefully evaluate the trade-off between the number of servers, the associated costs, and the potential impact on system efficiency and customer throughput.

Remark 2.7. *The choice of parameters in Table 2.5, while arbitrary, was carefully done to ensure the stability of the system and to observe clear behaviors that could be interpreted meaningfully. Based on the observations above,* *it is important to recognize that most of the results align with our intuition. However, there are some examples that are less straightforward to interpret, possibly due to the specific costs and parameters chosen.*

2.6 Conclusion

In this chapter, we presented a queueing system applicable to manufacturing systems producing electronic devices like smartphones, tablets, or laptops. Our model incorporated batch arrivals, multiple servers, catastrophic events, substitute service during breakdowns, customer balking and reneging behavior, and feedback. We derived the stability condition and employed probability generating functions to obtain closed-form expressions for the steady-state probabilities and performance measures. Furthermore, we conducted a numerical analysis to evaluate the impact of different parameters on key performance metrics, total expected cost, and total expected profit.

Potential future research directions include extending the proposed model to batch service queues. It would also be interesting to explore more complex scenarios, such as repairable queueing systems with non-Markovian arrival processes for customers and non-Markovian service processes for normal and breakdown services. Such extensions would enhance the model's applicability to a wider range of real-world manufacturing scenarios.

Chapter 3

Optimization analysis of unreliable multi-server queueing system with Bernoulli schedule working vacation, threshold-based recovery policy, and impatience

3.1 Introduction

With growth of communication systems and networks, manufacturing systems, transportation systems, etc, queueing systems with breakdowns have received growing significance [89, 121, 123].

Queueing models incorporating threshold policies, specifically the N-policy and F-policy, have garnered significant attention in recent years. The former policy dictates that a server activates only when N (where $N \geq 1$) or more customers accumulate in the system [68, 131, 39, 157]. Conversely, the latter policy restricts customer entry into the system once it reaches its capacity. When the queue length decreases to a threshold parameter value F, the server then permits customers to enter [53, 38, 81].

The literature on N and F policies is extensive. However, research on queueing models with breakdowns, repairs, and a threshold-based recovery policy, where the server remains unrepaired until the number of customers in the system reaches a predetermined threshold value, is limited. Notable works include [153, 63, 104].

Vacation queueing models have attracted substantial interest from researchers over the past decades, owing to their ubiquitous applications across diverse fields. These applications span production/manufacturing systems, telecommunication systems and computer networks. Notably, comprehensive surveys on this subject have been conducted by [48, 49, 135, 139].

The concept of working vacations was introduced by [119], proposing a model where the server processes jobs with varying intensities based on the incoming traffic. The primary objective is twofold: better control of queue lengths and reduction of customer loss. Additionally, working vacations enable servers to be strategically redirected for maintenance purposes. As a result, these models have gained significant popularity, leading to a wealth of analytical results in the literature, such as [159, 52, 155, 154].

In recent times, queueing systems that account for customer impatience have garnered significant attention. These models find realistic applications in various service systems and e-commerce domains. For a comprehensive overview of the literature on this theme, readers can refer to studies by [83, 23, 25, 21].

In this chapter, we delve into the analysis of a multi-server Markovian queue that integrates several crucial practical features including breakdowns, threshold-based recovery policy, working vacations, Bernoulli interruption schedule, impatient customers, and retention of reneged customers. The contributions and advantages of this paper are as follows:

1. ***The model.*** Unlike existing literature that predominantly focuses on single-server queueing models, our study embraces a multi-server queue. By incorporating the diverse features mentioned above, our proposed model offers greater flexibility in characterizing complex stochastic phenomena within multi-server machining systems.

2. ***Methodology and results.*** Leveraging the Q-matrix method, we provide a detailed theoretical analysis. We derive steady-state probabilities and various performance measures. Our chosen method is well-suited for analyzing quasi-birth-and-death (QBD) processes in steady-state.

3. ***Numerical illustrations.*** We develop a cost function to optimize service rates during both working vacation and normal busy periods. Additionally, we determine the optimal number of servers and explore threshold-based recovery policies. These insights empower system managers and decision-makers to regulate the system economically.

The chapter is structured concisely in the following manner: Section 2 presents the main motivation and practical applications of the current research work. Section 3 briefly describes the model under consideration. Section 4 comprises the analysis of the model in the stationary state. Section 5 enlists important performance measures. Section 6 develops a cost model for the proposed system and introduces cost optimization methods, namely, the direct search method and the quasi-Newton method. Section 7 deals with a cost optimization problem and provides numerical examples to illustrate the effects of different system parameters on performance measures, total expected cost, and total expected profit. Section 8 presents the conclusions of the study.

3.2 Main motivation and practical application

The motivating context for our model is analysis of automated teller machine (ATM) manufacturing systems. Such facilities commonly face machine failures and repairs, congestion issues, operator unavailability, impatient customers, and more that can significantly hamper production efficiency.

Specifically, we consider a production system with c parallel machines and finite finished goods capacity. Upon arrival of failed parts/sub-assemblies for repair, they immediately occupy any available operator. Otherwise failed units wait in queue for a random duration. Once all repairs are completed, operators take group vacations, relying on substitutes with slower service rates, and may have their breaks interrupted if failures resume.

Moreover, operators undergo their own failures following a breakdown process. Repairs only initiate after M failed machines have accumulated via a threshold policy. Newly arriving failures may balk from the repair queue or later renege after prolonged waits.

All such issues—breakdowns, vacations, congestion, balking and reneging—are commonly faced by real ATM manufacturers. By mathematically capturing these dynamics in a closed-form queueing model, we aim to evaluate the complex tradeoffs between maintainability, throughput, and customer impatience. The model can help optimize the number of machines, the threshold-based recovery policy, and service rates, to control costs in ATM production systems through resilience to inevitable disruptions.

3.3 Model description

Consider an Automated Manufacturing System modeled as an unreliable $M/M/c/L$ queueing system. The model formulation necessitates several distinct assumptions, which can be summarized as follows:

(i) Arrival process: Customers arrive following Poisson process with parameter α.

(ii) Service and working vacation processes:

  (a) Upon arrival, customers are served if any servers are available.

(b) After serving all existing customers, servers synchronously switch to a vacation period.

(c) Upon returning from vacation, if the system remains empty, servers immediately begin another synchronous vacation.

(d) The vacation duration follows an exponential distribution with parameter τ.

(e) During vacation, substitute servers take over from the main servers to serve new customers.

(f) Service times during regular busy periods (RBP) and vacations follow exponential distributions with parameters μ and ν, respectively. We assume that $\nu < \mu$.

(g) If a customer arrives and finds any of the c servers free (during busy or working vacation), they immediately occupy that server. If all servers are busy, the customer joins the end of the queue in the buffer and is served later according to the First-Come-First-Served (FCFS) discipline.

(iii) Bernoulli interruption scheme:

(a) During the working vacation period (WVP), the server operates under the Bernoulli rule. Specifically, at the instant of service completion during this period, if there are customers in the system:

- With probability β, the server interrupts the vacation and switches to the regular working period.
- With probability $\beta' = 1-\beta$, the server continues the vacation.

(b) Notably, the service during WVP is applied only to the first customer who arrives during this period.

Then, we can write

$$\delta_n = n\beta\nu\mathbb{1}_{2\leq n\leq c-1} + c\beta\nu\mathbb{1}_{c\leq L}.$$

(iv) Breakdown process: The system is susceptible to unreliability at any given time. During regular busy periods, servers are vulnerable to breakdowns. Specifically, a server break down only if there is at least one customer in the system. The occurrence of breakdowns follows a stationary Poisson process with parameter φ. Importantly, during repair periods (RP), customers cannot be served.

(v) The threshold-based recovery policy and repair process: The recovery can be performed when M $(1 \leq M \leq c-1)$ or more customers are present. The repair period has exponential distribution with parameter γ. Customers arriving during the repair time are ignored by the system.

(vi) Balking: When a customer arrives, their actions depend on server availability:

- - If some servers are working and others are free, the customer is directly served.
- - Otherwise, during working vacation, regular busy, or repair periods:
  (a) The customer may join the queue with probability θ_n.
  (b) Customers faced with joining a queue have an alternative: they may balk, choosing not to enter. The balking probability is denoted as: $\theta'_n = 1 - \theta_n$, where in the case of working vacation/regular busy period, we have : $0 \leq \theta_{n+1} \leq \theta_n \leq 1$. Consider the following scenarios:
    i. For working vacation/regular busy period case, we have:
      – $0 \leq \theta_{n+1} \leq \theta_n \leq 1$ for $c \leq n \leq L-1$;
      – $\theta_0 = 1, \ldots, \theta_{c-1} = 1$.
    ii. For repair period, we observe:
      – $0 \leq \theta_{n+1} \leq \theta_n \leq 1$ for $1 \leq n \leq L-1$.
      – $\theta_0 = 1$ (no balking when the system is empty).
    iii. In both cases, we have: $\theta_L = 0$ (no entering when the system is at full capacity).

Shortly, we have for working vacation and regular normal busy:

$$\alpha_n = \alpha \mathbb{1}_{n<c} + \theta_n \alpha \mathbb{1}_{c \leq n \leq L},$$

and for breakdown period:

$$\alpha_n = \quad \theta_n \alpha, \quad 1 \leq n \leq L.$$

(vii) Reneging and retention:

(a) Upon arrival, customers exhibit different behaviors based on the server status:
  - If servers are in regular working mode or working vacation period:
    – The customer activates an impatience timer T_1 (for regular working) or T_0 (for working vacation). If the customer's service is not completed before the timer expires, they may abandon the system.
  - During the reparation period:
    – A new arrival activates its own timer T_2. If service is unavailable before the expiration of the impatience timer, the customer may give up.

(b) The impatience time T_j follows an exponentially distributed random variable with rates $\varsigma_j > 0$ (where $j = 0, 1, 2$).

(c) Impatient customers have two options:

- They may quit the system without receiving service with probability κ.
- Alternatively, they may be kept in the system with probability $\kappa' = 1 - \kappa$.

Then, we can put:

$$\epsilon_{n,j} = n\kappa\varsigma_0 \mathbb{1}_{j=0} + n\kappa\varsigma_1 \mathbb{1}_{j=1} + n\kappa\varsigma_2 \mathbb{1}_{j=2},$$

$$\upsilon_n = (n\mu + \epsilon_{n,1})\mathbb{1}_{1\leq n\leq c-1} + (c\mu + \epsilon_{n,1})\mathbb{1}_{c\leq n\leq L},$$

and

$$\zeta_n = (\nu + \epsilon_{1,0})\mathbb{1}_{n=1} + (n\beta'\nu + \epsilon_{n,0})\mathbb{1}_{2\leq n\leq c-1} + (c\beta'\nu + \epsilon_{n,0})\mathbb{1}_{c\leq n\leq L}.$$

The customers timers are independent and identically distributed (i.i.d.) random variables and independent of the number of customers currently waiting.

(viii) The various stochastic processes within the system are assumed to be mutually independent.

3.4 Equilibrium probability analysis

We employ the Markov process approach, utilizing the Q-matrix, to establish the steady-state distribution for our proposed queueing model. Our primary focus lies in deriving the steady-state probabilities of the system, specifically as a function of the probability $\pi_{1,1}$, rather than relying on $\pi_{0,j}$ or $\pi_{L,j}$ for $j = 0, 1, 2$.

The system under consideration can be modeled as a continuous-time Markov process, denoted by $\{\mathfrak{X}(t), \mathfrak{Y}(t); t \geq 0\}$, where $\mathfrak{X}(t)$ represents the number of customers present in the system at time t, and $\mathfrak{Y}(t)$ characterizes the operational state of the servers at time t. The possible states for $\mathfrak{Y}(t)$ are as follows:

$$\mathfrak{Y}(t) = \begin{cases} 0, & \text{Servers are in a WVP} \\ 1, & \text{Servers are in a RBP} \\ 2, & \text{Servers are in a RP} \end{cases}$$

Let $\pi_{n,j}$ denote the steady-state probability that the system has n customers and the servers are in state j, such that: $\pi_{n,j} = \lim_{t\to\infty} P\{\mathfrak{X}(t) = n, \mathfrak{Y}(t) = j\}$, where

$(n, j) \in \{\{(n, 0) : n = 0, 1, ..., L\} \cup \{(n, 1) : n = 1, 2, ..., L\} \cup \{(n, 2) : n = 1, 2, ..., L\}\}$. The state transition rate diagram is depicted in Figure 3.1.

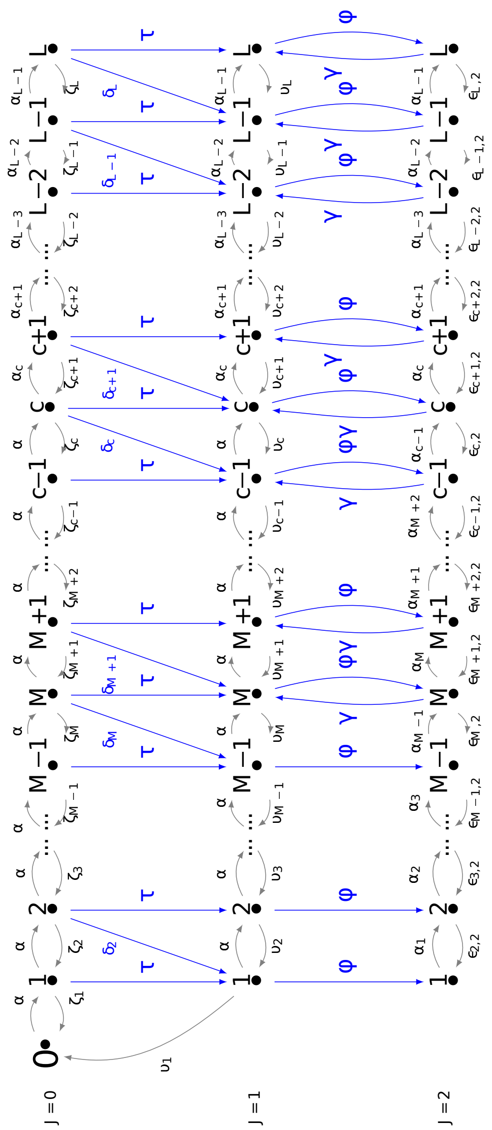

Figure 3.1: State transition diagram for the proposed model

3.4.1 Governing equations

The steady-state balance equations that govern our system are expressed as follows:

$$
\begin{aligned}
\alpha\pi_{0,0} &= (\mu+\kappa\varsigma_1)\pi_{1,1}+(\nu+\kappa\varsigma_0)\pi_{1,0}, \;\; n=0,\\
(\alpha+n(\nu+\kappa\varsigma_0)+\tau)\pi_{n,0} &= \alpha\pi_{n-1,0}+(n+1)(\beta'\nu+\kappa\varsigma_0)\pi_{n+1,0},\\
&\quad 1\le n\le c-1,\\
(\alpha\theta_n+c\nu+n\kappa\varsigma_0+\tau)\pi_{n,0} &= \alpha\pi_{n-1,0}+(c\beta'\nu+(n+1)\kappa\varsigma_0)\pi_{n+1,0},\\
&\quad n=c,\\
(\alpha\theta_n+c\nu+n\kappa\varsigma_0+\tau)\pi_{n,0} &= \alpha\theta_{n-1}\pi_{n-1,0}+(c\beta'\nu+(n+1)\kappa\varsigma_0)\pi_{n+1,0},\\
&\quad c+1\le n\le L-1,\\
(\tau+c\nu+L\kappa\varsigma_0)\pi_{L,0} &= \alpha\theta_{L-1}\pi_{n-1,0}, \quad n=L,\\
(\alpha+\mu+\kappa\varsigma_1+\varphi)\pi_{1,1} &= \tau\pi_{1,0}+2\beta\nu\pi_{2,0}+2(\mu+\kappa\varsigma_1)\pi_{2,1}, \quad n=1,\\
(\alpha+n(\mu+\kappa\varsigma_1)+\varphi)\pi_{n,1} &= \alpha\pi_{n-1,1}+(n+1)\beta\nu\pi_{n+1,0}+(n+1)(\mu+\kappa\varsigma_1)\pi_{n+1,1}\\
&\quad +\tau\pi_{n,0}, \;\; 2\le n\le M-1,\\
(\alpha+n(\mu+\kappa\varsigma_1)+\varphi)\pi_{n,1} &= \alpha\pi_{n-1,1}+(n+1)\beta\nu\pi_{n+1,0}+(n+1)(\mu+\kappa\varsigma_1)\pi_{n+1,1}\\
&\quad +\tau\pi_{n,0}+\gamma\pi_{n,2}, \quad M\le n\le c-1,\\
(\theta_n\alpha+n(\mu+\kappa\varsigma_1)+\varphi)\pi_{n,1} &= \alpha\pi_{n-1,1}+c\beta\nu\pi_{n+1,0}+(c\mu+(n+1)\kappa\varsigma_1)\pi_{n+1,1}\\
&\quad +\tau\pi_{n,0}+\gamma\pi_{n,2}, \quad n=c,\\
(\alpha\theta_n+n(\mu+\kappa\varsigma_1)+\varphi)\pi_{n,1} &= \alpha\theta_{n-1}\pi_{n-1,1}+c\beta\nu\pi_{n+1,0}+(c\mu+(n+1)\kappa\varsigma_1)\pi_{n+1,1}\\
&\quad +\tau\pi_{n,0}+\gamma\pi_{n,2}, \quad c+1\le n\le L-1,\\
(c\mu+L\kappa\varsigma_1+\varphi)\pi_{L,1} &= \alpha\theta_{L-1}\pi_{L-1,1}+\tau\pi_{L,0}+\gamma\pi_{L,2}, \quad n=L,\\
\theta_1\alpha\pi_{1,2} &= \varphi\pi_{1,1}+2\kappa\varsigma_2\pi_{2,2}, \quad n=1,\\
(\alpha\theta_n+n\kappa\varsigma_2)\pi_{n,2} &= \alpha\theta_{n-1}\pi_{n-1,2}+(n+1)\kappa\varsigma_2\pi_{n+1,2}+\varphi\pi_{n,1},\\
&\quad 2\le n\le M-1,\\
(\alpha\theta_n+n\kappa\varsigma_2+\gamma)\pi_{n,2} &= \alpha\theta_{n-1}\pi_{n-1,2}+(n+1)\kappa\varsigma_2\pi_{n+1,2}+\varphi\pi_{n,1},\\
&\quad M\le n\le L-1,\\
(L\kappa\varsigma_2+\gamma)\pi_{L,2} &= \alpha\theta_{L-1}\pi_{L-1,2}+\varphi\pi_{L,1}, \quad n=L.
\end{aligned}
$$

The normalizing condition is expressed as:

$$\sum_{n=0}^{L}\pi_{n,0}+\sum_{n=1}^{L}\pi_{n,1}+\sum_{n=1}^{L}\pi_{n,2}=1. \tag{3.1}$$

Let's introduce the necessary notations for the subsequent sections of the paper:

$$\dot{\zeta}_n=\begin{cases} n\nu+\epsilon_{n,0}, & 1\le n\le c-1,\\ c\nu+\epsilon_{n,0}, & c\le n\le L,\end{cases}$$

$$
\varrho_n = \begin{cases} -(\alpha + \upsilon_n + \varphi), & 1 \leq n \leq c-1, \\ -(\alpha_c + \upsilon_n + \varphi), & n = c, \\ -(\alpha_n + \upsilon_n + \varphi), & c+1 \leq n \leq L-1, \\ -(\upsilon_L + \varphi), & n = L, \end{cases}
$$

$$
\vartheta_n = \begin{cases} -(\alpha + \dot{\zeta}_n + \tau), & 1 \leq n \leq c-1, \\ -(\alpha_c + \dot{\zeta}_n + \tau), & n = c, \\ -(\alpha_n + \dot{\zeta}_n + \tau), & c+1 \leq n \leq L-1, \\ -(\dot{\zeta}_L + \tau), & n = L, \end{cases}
$$

$$
\sigma_n = \begin{cases} -\alpha_1, & n = 1, \\ -(\alpha_n + \epsilon_{n,2}), & 2 \leq n \leq M-1, \\ -(\alpha_n + \epsilon_{n,2} + \gamma), & M \leq n \leq L-1, \\ -(\epsilon_{L,2} + \gamma), & n = L. \end{cases}
$$

3.4.2 Matrix solution

To obtain the steady-state solution, let the steady-state probability vector of the infinitesimal generator Q be denoted as $P = (\pi_0, \pi_1, \pi_2)$, where $\pi_0 = (\pi_{0,0}, \pi_{1,0}, ..., \pi_{L,0})$, $\pi_1 = (\pi_{1,1}, \pi_{2,1}, ..., \pi_{L,1})$, and $\pi_2 = (\pi_{1,2}, \pi_{2,2}, ..., \pi_{L,2})$. The steady-state equations $PQ = 0$ must be satisfied by P along with the normalization condition $Pe = 1$, where 0 is a zero row vector, and $e = (e_0, e_1, e_2)$ is a $(3L+1)$ column vector of ones, with e_0 being an $(L+1)$ column vector and e_1 and e_2 being L column vectors. The block structure of the infinitesimal generator Q is as follows:

$$
Q = \begin{pmatrix} \mathcal{A}_1 & \mathcal{A}_2 & \mathcal{A}_3 \\ \mathcal{E}_1 & \mathcal{E}_2 & \mathcal{E}_3 \\ \mathcal{D}_1 & \mathcal{D}_2 & \mathcal{D}_3 \end{pmatrix},
$$

where

$$
\mathcal{A}_1 = \begin{pmatrix} -\alpha & \alpha & & & & & & & & \\ \zeta_1 & \vartheta_1 & \alpha & & & & & & & \\ & \zeta_2 & \vartheta_2 & \alpha & & & & & & \\ & & \ddots & \ddots & \ddots & & & & & \\ & & & \zeta_{c-1} & \vartheta_{c-1} & \alpha & & & & \\ & & & & \zeta_c & \vartheta_c & \alpha_c & & & \\ & & & & & \ddots & \ddots & \ddots & & \\ & & & & & & \zeta_{L-1} & \vartheta_{L-1} & \alpha_{L-1} \\ & & & & & & 0 & \zeta_L & \vartheta_L \end{pmatrix},
$$

$$
\mathcal{A}_2 = \begin{pmatrix} 0 & 0 & ... & ... & 0 \\ \tau & 0 & ... & ... & 0 \\ 0 & \tau & ... & ... & 0 \\ \vdots & \vdots & & & \vdots \\ 0 & 0 & ... & ... & \tau \end{pmatrix},
$$

$$\mathcal{E}_2 = \begin{pmatrix} \varrho_1 & \alpha & & & & & & & \\ \upsilon_2 & \varrho_2 & \alpha & & & & & & \\ & \upsilon_3 & \varrho_3 & \alpha & & & & & \\ & & \ddots & \ddots & \ddots & & & & \\ & & & \upsilon_{c-1} & \varrho_{c-1} & \alpha & & & \\ & & & & \upsilon_c & \varrho_c & \alpha_c & & \\ & & & & & \ddots & \ddots & \ddots & \\ & & & & & & \upsilon_{L-1} & \varrho_{L-1} & \alpha_{L-1} \\ & & & & & & 0 & \upsilon_L & \varrho_L \end{pmatrix},$$

$$\mathcal{D}_3 = \begin{pmatrix} \sigma_1 & \alpha_1 & & & & \\ \epsilon_{2,2} & \sigma_2 & \alpha_2 & & & \\ & \epsilon_{3,2} & \sigma_3 & \alpha_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & \epsilon_{L-1,2} & \sigma_{L-1} & \alpha_{L-1} \\ & & & 0 & \epsilon_{L,2} & \sigma_L \end{pmatrix},$$

$$\mathcal{E}_1 = \begin{bmatrix} \upsilon_1 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix}, \quad \mathcal{E}_3 = \begin{bmatrix} \varphi & & \\ & \ddots & \\ & & \varphi \end{bmatrix}, \quad \text{and } \mathcal{D}_2 = \begin{bmatrix} 0 & & & & & \\ & \ddots & & & & \\ & & 0 & & & \\ & & & \gamma & & \\ & & & & \ddots & \\ & & & & & \gamma \end{bmatrix}.$$

Note that $\mathcal{A}_1$ and $\mathcal{A}_2$ are matrices of dimensions $(L+1) \times (L+1)$ and $(L+1) \times L$, respectively. $\mathcal{A}_3$ is a zero matrix with dimensions $(L+1) \times L$. $\mathcal{E}_1$ is an $L \times (L+1)$ matrix. $\mathcal{E}_2$, $\mathcal{E}_3$, $\mathcal{D}_2$, and $\mathcal{D}_3$ are square matrices of order $L \times L$. Additionally, $\mathcal{D}_1$ is a zero matrix with dimensions $L \times (L+1)$.

3.4.3 System state probabilities

We present the steady state equations $PQ = 0$ and the normalization condition $Pe = 1$, as

$$\begin{cases} \pi_0\mathcal{A}_1 + \pi_1\mathcal{E}_1 + \pi_2\mathcal{D}_1 = 0, \\ \pi_0\mathcal{A}_2 + \pi_1\mathcal{E}_2 + \pi_2\mathcal{D}_2 = 0, \\ \pi_0\mathcal{A}_3 + \pi_1\mathcal{E}_3 + \pi_2\mathcal{D}_3 = 0, \\ \pi_0 e_1 + \pi_1 e_2 + \pi_2 e_3 = 1. \end{cases} \tag{3.2}$$

As $\mathcal{A}_3$ and $\mathcal{D}_1$ are null matrices, then Equation (3.2) can be rewritten as

$$\pi_0\mathcal{A}_1 + \pi_1\mathcal{E}_1 = 0, \tag{3.3}$$

$$\pi_0\mathcal{A}_2 + \pi_1\mathcal{E}_2 + \pi_2\mathcal{D}_2 = 0, \tag{3.4}$$

$$\pi_1\mathcal{E}_3 + \pi_2\mathcal{D}_3 = 0, \tag{3.5}$$

$$\pi_0 e_1 + \pi_1 e_2 + \pi_2 e_3 = 1. \tag{3.6}$$

Next, put $\mathcal{A}_2 = \begin{pmatrix} O_1 \\ \tau I_L \end{pmatrix}$, $\mathcal{E}_1 = \begin{pmatrix} \upsilon_1 & O_1 \\ O_2 & O_3 \end{pmatrix}$, $\mathcal{E}_3 = (\varphi I_n)$, $\mathcal{D}_2 = \begin{pmatrix} O_4 & \\ \dot{O}_4 & \gamma I_{L-M+1} \end{pmatrix}$, with I_L denotes the identity matrix. Further, O_1 is a $1 \times L$ matrix. O_2 and O_3 are both of order $(L-1) \times 1$ and $(L-1) \times L$, respectively. O_4 has dimensions $(M-1) \times L$. $\dot{O}_4$ is of order $(L-M+1) \times (M-1)$. I_{L-M+1} represents the identity matrix of order $L-M+1$.

Let $\mathcal{A}1^{-1}$ and $\mathcal{D}3^{-1}$ denote the inverse matrices of $\mathcal{A}1$ and $\mathcal{D}3$, respectively. By referring to Eq. (3.3), we obtain the following result:

$$\begin{aligned} \pi_0 &= -\pi_1 \mathcal{E}_1 \mathcal{A}_1^{-1} \\ &= -\pi_1 \begin{pmatrix} \upsilon_1 o \\ O_5 \end{pmatrix} \\ &= -\pi_{1,1} \upsilon_1 o, \end{aligned} \tag{3.7}$$

where $o = (o_0, \tilde{o})$, such that $\tilde{o} = (o_1, ..., o_L)$ be an L row vector of the matrix $\mathcal{A}_1^{-1}$, and O_5 is $(L-1) \times (L+1)$. From Eq. (3.5), we have

$$\pi_2 = -\pi_1 \mathcal{E}_3 \mathcal{D}_3^{-1} = -\pi_1 \varphi \mathcal{D}_3^{-1}. \tag{3.8}$$

Substituting Eqs. (3.7) and (3.8) into Eq. (3.4), obtain

$$-\pi_{1,1} \upsilon_1 \tilde{o} \tau + \pi_1 (\mathcal{E}_2 - \varphi \mathcal{D}_3^{-1} \mathcal{D}_2) = 0. \tag{3.9}$$

As $\mathcal{E}_2$ and $\mathcal{D}_3^{-1}$ are both square matrices of order L, we can affirm the existence of the matrix:

$$\tilde{\mathcal{E}} = (\mathcal{E}_2 - \varphi \mathcal{D}_3^{-1} \mathcal{D}_2)^{-1}.$$

Thus

$$\pi_1 = (\upsilon_1 \tilde{o} \tau \tilde{\mathcal{E}}) \pi_{1,1}. \tag{3.10}$$

Consequently, we can deduce easily

$$\pi_2 = -(\upsilon_1 \tilde{o} \tau \tilde{\mathcal{E}} \varphi \mathcal{D}_3^{-1}) \pi_{1,1}. \tag{3.11}$$

Then, using Eqs. (3.7)–(3.11), we get:

$$\begin{cases} \pi_{n,0} = -\upsilon_1 o_n \pi_{1,1}, \\ \pi_{n,1} = (\upsilon_1 \tau \sum_{i=1}^{L} o_{i+1} \tilde{\psi}_{in}) \pi_{1,1}, \\ \pi_{n,2} = -\left(\upsilon_1 \tau \varphi \sum_{j=1}^{L} \sum_{i=1}^{L} o_{i+1} \tilde{\psi}_{ij} \tilde{\omega}_{jn} \right) \pi_{1,1}, \end{cases}$$

where $\tilde{\omega}_{ij}$ are the elements of matrix $\tilde{\mathcal{D}} = \mathcal{D}_3^{-1}$, and $\tilde{\psi}_{ij}$ are the elements of the matrix $\tilde{\mathcal{E}} = (\mathcal{E}_2 - \varphi \mathcal{D}_3^{-1} \mathcal{D}_2)^{-1}$. Finally, to determine $\pi_{1,1}$, we apply the normalizing condition (as described in Equation (3.1)):

$$\pi_{1,1} = \left(-\upsilon_1 \sum_{n=0}^{L} o_n + \upsilon_1 \tau \sum_{n=1}^{L} \sum_{i=1}^{L} o_{i+1} \tilde{\psi}_{in} - \upsilon_1 \tau \varphi \sum_{n=1}^{L} \sum_{j=1}^{L} \sum_{i=1}^{L} o_{i+1} \tilde{\psi}_{ij} \tilde{\omega}_{jn} \right)^{-1}.$$

3.5 Performance measures

In this section, we delve into the derivation of crucial system indices, leveraging the probabilities associated with the system distribution.

Result 1: The servers are in busy period with probability

$$P_{busy} = \sum_{n=1}^{L} \pi_{n,1} = \upsilon_1 \tau \sum_{n=1}^{L} \sum_{i=1}^{L} o_{i+1} \tilde{\psi}_{in} \pi_{1,1}. \tag{3.12}$$

Result 2: The servers are in working vacation period with probability

$$P_{wv} = \sum_{n=1}^{L} \pi_{n,0} = -\left(\upsilon_1 \sum_{n=0}^{L} o_n \right) \pi_{1,1}. \tag{3.13}$$

Result 3: The servers are in breakdown period with probability

$$P_{bp} = \sum_{n=1}^{L} \pi_{n,2} = -\upsilon_1 \tau \varphi \left(\sum_{n=1}^{L} \sum_{j=1}^{L} \sum_{i=1}^{L} o_{i+1} \tilde{\psi}_{ij} \tilde{\omega}_{jn} \right) \pi_{1,1}. \tag{3.14}$$

Result 3: The probability of system reliability

$$P_{re} = 1 - \pi_{pb}.$$

Result 4: The mean system size is

$$\begin{aligned} E_s &= \sum_{n=1}^{L} n(\pi_{n,0} + \pi_{n,1} + \pi_{n,2}) \qquad (3.15) \\ &= \upsilon_1 \left(-\sum_{n=1}^{L} n o_n + \tau \sum_{n=1}^{L} \sum_{i=1}^{L} n o_{i+1} \tilde{\psi}_{in} - \tau \varphi \sum_{n=1}^{L} \sum_{j=1}^{L} \sum_{i=1}^{L} n o_{i+1} \tilde{\psi}_{ij} \tilde{\omega}_{jn} \right) \pi_{1,1}. \end{aligned}$$

Result 5: The effective arrival rate

$$\alpha' = \alpha \pi_{0,0} + \sum_{n=1}^{L} \alpha_n \pi_{n,0} + \sum_{n=1}^{L} \alpha_n \pi_{n,1} + \sum_{n=1}^{L} \alpha_n \pi_{n,2}.$$

Result 6: The mean waiting time of customers in the system

$$W_s = E_s / \alpha^{'}.$$

Result 7: The average balking rate

$$R_{balk} = \alpha - \alpha^{'}. \tag{3.16}$$

Result 8: The average reneging rate

$$\begin{aligned} R_{ren} &= \kappa\varsigma_0 \sum_{n=1}^{L} n\pi_{n,0} + \kappa\varsigma_1 \sum_{n=1}^{L} n\pi_{n,1} + \kappa\varsigma_2 \sum_{n=2}^{L} n\pi_{n,2} \qquad (3.17) \\ &= \upsilon_1 \kappa \left(-\varsigma_0 \sum_{n=1}^{L} no_n + \tau\varsigma_1 \sum_{n=1}^{L}\sum_{i=1}^{L} no_{i+1}\tilde{\psi}_{in} - \tau\varphi\varsigma_2 \sum_{n=2}^{L}\sum_{j=1}^{L}\sum_{i=1}^{L} no_{i+1}\tilde{\psi}_{ij}\tilde{\omega}_{jn} \right) \pi_{1,1}. \end{aligned}$$

Result 9: The average retention rate

$$\begin{aligned} R_{ret} &= \kappa^{'}\varsigma_0 \sum_{n=1}^{L} n\pi_{n,0} + \kappa^{'}\varsigma_1 \sum_{n=1}^{L} n\pi_{n,1} + \kappa^{'}\varsigma_2 \sum_{n=2}^{L} n\pi_{n,2} \qquad (3.18) \\ &= \upsilon_1 \kappa^{'} \left(-\varsigma_0 \sum_{n=1}^{L} no_n + \tau\varsigma_1 \sum_{n=1}^{L}\sum_{i=1}^{L} no_{i+1}\tilde{\psi}_{in} - \tau\varphi\varsigma_2 \sum_{n=2}^{L}\sum_{j=1}^{L}\sum_{i=1}^{L} no_{i+1}\tilde{\psi}_{ij}\tilde{\omega}_{jn} \right) \pi_{1,1}. \end{aligned}$$

Result 10: The mean number of customers served per unit time

$$C_s = \nu \sum_{n=1}^{L} n\pi_{n,0} + \mu \sum_{n=1}^{L} n\pi_{n,1}.$$

3.6 Cost model and optimization

For our queueing model, we consider he cost components as outlined below:

1. C_{busy}: unit time cost for system being in busy period.
2. C_{wv}: unit time cost for system being is in working vacation.
3. C_{break}: unit time cost for system being in breakdown period.
4. C_{sq}: Holding unit time cost when a customer enters the queue.
5. C_{s_1}: Cost per service per unit time in regular working period.
6. C_{s_2}: Cost per service per unit time in working vacation period.
7. C_l: unit time cost when a customer is lost.
8. C_t: unit time cost when the system retains a customer.
9. C_f: Fixed purchase cost of the server per unit.

The formulation of the cost per unit time function for the queueing system is as follows:

$$\begin{aligned} \mathcal{T}_c &= C_{busy}P_{busy} + C_{wv}P_{wv} + C_{break}P_{pb} + C_{sq}E_s + C_l(R_{ren} + R_{balk}) \qquad (3.19) \\ &\quad + C_tR_{ret} + c(\mu C_{s_1} + \nu C_{s_2}) + cC_f. \end{aligned}$$

Expressing the expected cost function $\mathcal{T}_c$ explicitly by substituting Equations (3.12)–(3.18) into (3.19) would result in an extremely complex formulation. Consequently, studying the analytical behavior of $\mathcal{T}_c$ becomes a big challenge. Furthermore, due to the nonlinearity and intricacy of the expected cost function, deriving the optimal solution (c^*, M^*, μ^*, ν^*) in closed form would be an arduous task.

To circumvent these difficulties and perform the optimization analysis, we employ direct search and Newton's methods as numerical optimization techniques to search for the optimal solution (c^*, M^*, μ^*, ν^*). Initially, the direct search method is utilized to determine the optimal values of the variables (c^*, M^*). Subsequently, with these variables fixed, Newton's method is applied to find the optimal values of the variables (μ^*, ν^*).

3.6.1 Numerical cost optimum parameter

We consider a practical problem concerning the automated teller machine (ATM) production facility mentioned in Section 2. In the context of the considered practical example, the system parameters are delineated as follows: failed machines arrive according to a Poisson process with $\alpha = 7$. The system capacity is considered finite with $L = 20$. If the system is in operation, the failure occurs in which the breakdown times are exponential distribution with $\varphi = 0.1$. The service and repair times of the machines obey exponential distributions with parameters $\mu = 3.0$, $\nu = 0.9$, and $\gamma = 0.3$, respectively. Once the system gets empty, it goes on vacation period, the vacation period follows exceptional distribution with parameter $\tau = 0.4$. The failed machines during both period may get impatient and leave the system with being served. The impatience timers follow exponential distribution with $\varsigma_0 = 0.5$, $\varsigma_1 = 0.3$, $\varsigma_2 = 0.9$. Further, during working vacation period, the failed machines service may be continue their service during working vacation period with probability $\beta' = 0.6$, and they can leave the system with probability $\kappa = 0.7$. The joining probability is taken as $\theta_n = 1 - \frac{n}{L}$.

An efficient algorithm based on the direct search method is employed to determine the optimal discrete values (c^*, M^*) that optimize the expected cost function. The effectiveness of this approach hinges on the convexity (or unimodality) of the cost function. Throughout the numerical analysis, the following cost elements are considered: $C_{busy} = \$20$, $C_{wv} = \$20$, $C_{bp} = \$50$, $C_{sq} = \$10$, $C_{s1} = \$5$, $C_{s2} = \$5$, $C_l = \$30$, $C_t = \$25$, $C_f = \$1$ and $R = 50$.

Figure 3.2 illustrates the behavior of the expected cost function $\mathcal{T}_c(M^*, c^*)$ for varying values of c and M. The plotted curve exhibits a convex shape, indicating the existence of a single relative minimum. Consulting Table 3.1, it is evident that the minimum expected cost per unit time, which amounts to **182.5710**, is attained when $M^* = 6$ and $c^* = 1$.

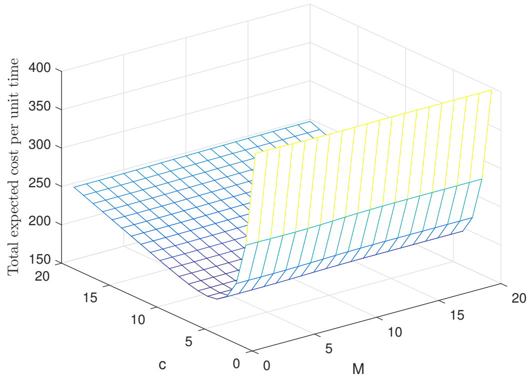

Figure 3.2: The expected cost $\mathcal{T}_c$ for different values of c and M.

Table 3.1: c and M vs. $\mathcal{T}_c(M, c)$

M / c	1	2	3	4	5	6	7	8
2	271.9823	-	-	-	-	-	-	-
3	216.2269	216.2388	-	-	-	-	-	-
4	193.3445	193.3762	193.3707	-	-	-	-	-
5	184.5919	184.6346	184.6462	184.7100	-	-	-	-
6	**182.5710**	182.6202	182.6410	182.7123	182.8470	-	-	-
7	183.9711	184.0241	184.0501	184.1262	184.2599	184.4512	-	-
8	187.1491	187.2043	187.2331	187.3123	187.4466	187.6304	187.8610	-
9	191.2406	191.2970	191.3274	191.4083	191.5437	191.7250	191.9432	192.1971

Once the optimal values (M^*, c^*) are determined, Newton's method is employed to locate the minimum value of $\mathcal{T}_c(M^*, c^*, \mu^*, \nu^*)$ by iteratively optimizing the continuous variables μ and ν. Newton's method is an efficient iterative technique for finding the optimum of a nonlinear function by computing the search direction at each iteration.

The Quasi-Newton method, a variant of Newton's method, is utilized to numerically determine μ^* and ν^*. Numerical results obtained through this optimization process are presented in Tables 3.2-3.7 for various system parameter settings.

Table 3.2: $\mathcal{T}_c(c^*, M^*, \mu^*, \nu^*)$ and (c^*, M^*, μ^*, ν^*) while adjusting τ and α ($\varsigma_0 = 0.5$, $\varsigma_1 = 0.3$, $\varsigma_2 = 0.9, L = 20$, $\kappa = 0.7$, $\beta' = 0.6$, $\varphi = 0.1$, $\gamma = 0.3$)

(τ , α)	(2,5)	(2,8)	(2.5,5)	(2.5,8)	(3.0,5)	(3.0,8)
(c^* ,M^*)	(1,2)	(4,7)	(1,2)	(4,7)	(1,2)	(4,7)
μ^*	3.3953	1.4294	3.5066	1.4697	3.5893	1.4991
ν^*	1.7352	0.4679	1.3760	0.3242	1.0157	0.1853
$\mathcal{T}_c(c^*, M^*, \mu^*, \nu^*)$	125.1117	182.8660	124.4980	180.6144	123.3919	177.7679

Table 3.3: $\mathcal{T}_c(c^*, M^*, \mu^*, \nu^*)$ and (c^*, M^*, μ^*, ν^*) while adjusting β' ($\varsigma_0 = 0.5$, $\varsigma_1 = 0.3$, $\varsigma_2 = 0.9$,$L = 20$, $\kappa = 0.7$, $\alpha = 7$, $\varphi = 0.1$, $\tau = 0.4$, $\gamma = 0.3$)

β'	0.75	0.8	0.85	0.9	0.95
(c^* , M^*)	(3,4)	(4,5)	(4,6)	(5,7)	(6,8)
μ^*	1.6362	1.3007	1.0810	0.9233	0.8065
ν^*	1.5967	1.2149	0.9672	0.7970	0.6748
$\mathcal{T}_c(c^*, M^*, \mu^*, \nu^*)$	147.4804	150.3079	153.4443	156.8410	160.5053

Table 3.4: $\mathcal{T}_c(c^*, M^*, \mu^*, \nu^*)$ and (c^*, M^*, μ^*, ν^*) while adjusting L and κ ($\varsigma_0 = 0.5$, $\varsigma_1 = 0.3$, $\varsigma_2 = 0.9$, $\alpha = 7$, $\tau = 0.4$, $\beta' = 0.6$, $\varphi = 0.1$ $\gamma = 0.3$)

(L,κ)	(10,0.6)	(40,0.6)	(10,0.7)	(40,0.7)	(10,0.9)	(40,0.9)
(c^* ,M^*)	(7,1)	(7,4)	(7,1)	(7,4)	(7,1)	(7,4)
μ^*	0.9190	1.0397	0.8876	0.9821	0.8333	0.8757
ν^*	0.9180	0.8733	0.8866	0.8594	0.8323	0.8333
$\mathcal{T}_c(c^*, M^*, \mu^*, \nu^*)$	167.7709	167.6681	165.5115	165.0619	161.7871	160.6461

Table 3.5: $\mathcal{T}_c(c^*, M^*, \mu^*, \nu^*)$ and (c^*, M^*, μ^*, ν^*) while adjusting φ and γ ($\varsigma_0 = 0.5$, $\varsigma_1 = 0.3$, $\varsigma_2 = 0.9$, $\kappa = 0.7$,$\beta' = 0.6$, $\tau = 0.4$, $\alpha = 7$,$L = 20$)

(φ , γ)	(0.4 1)	(0.4 5)	(0.6 1)	(0.6 5)	(0.8 1)	(0.8 5)
μ^*	0.9571	0.9275	0.9299	0.8955	0.9691	0.9342
ν^*	0.8955	0.8907	0.9178	0.8945	0.9075	0.9030
$\mathcal{T}_c(c^*, M^*, \mu^*, \nu^*)$	170.3847	162.9654	172.9348	163.4767	177.8815	165.2887

Table 3.6: $\mathcal{T}_c(M^*, c^*, \mu^*, \nu^*)$ and (M^*, c^*, μ^*, ν^*) while adjusting ς_0 , ς_1 and $\varsigma_2 = 1.9$ ($\kappa = 0.7$,$L = 20$,$\alpha = 7$, $\tau = 0.4$,$\varphi = 0.1$, $\beta' = 0.6$,$\gamma = 0.3$)

(ς_1 ς_0)	(0.2,0.6)	(0.4,0.6)	(0.2,0.8)	(0.4,0.8)	(0.2,1.0)	(0.4,1.0)
(M^* ,c^*)	(8,2)	(8,4)	(8,2)	(8,3)	(8,5)	(8,7)
μ^*	0.8287	0.7616	0.8079	0.7640	0.8041	0.7644
ν^*	0.7697	0.7606	0.8056	0.7630	0.8031	0.7634
$\mathcal{T}_c(M^*, c^*, \mu^*, \nu^*)$	171.4859	172.4115	179.1606	180.3308	186.5172	188.2080

Table 3.7: $\mathcal{T}_c(M^*, c^*, \mu^*, \nu^*)$ and (M^*, c^*, μ^*, ν^*) while adjusting ς_0 , ς_1 and $\varsigma_2 = 3.1$ ($\kappa = 0.7$,$L = 20$,$\alpha = 7$, $\tau = 0.4$,$\varphi = 0.1$, $\beta' = 0.6$,$\gamma = 0.3$)

(ς_1 ς_0)	(0.2,0.6)	(0.4,0.6)	(0.2,0.8)	(0.4,0.8)	(0.2,1.0)	(0.4,1.0)
(M^* ,c^*)	(8,2)	(8,4)	(8,2)	(8,3)	(8,5)	(8,7)
μ^*	0.8211	0.7583	0.8032	0.7607	0.8006	0.7621
ν^*	0.7712	0.7573	0.8022	0.7597	0.7996	0.7611
$\mathcal{T}_c(M^*, c^*, \mu^*, \nu^*)$	170.7178	171.7406	178.4237	179.6992	185.8408	187.5095

Tables 3.2-3.7 illustrate the relationships between various system parameters and the optimal service rates (μ^*, ν^*) that minimize the expected cost $\mathcal{T}_c(M^*, c^*, \mu^*, \nu^*)$:

- As the arrival rate of failed machines (α) increases, the expected cost $\mathcal{T}_c(M^*, c^*, \mu^*, \nu^*)$ rises substantially. Similarity for β'. This is understandable, as a higher influx of failures naturally strains the service system, leading to longer queues, more congestion, and ultimately increased costs.
- Conversely, higher operator vacation rate (τ) and greater customer non-retention probability (κ) decrease the expected cost. Obviously, more frequent vacations provide more opportunities to serve customers during vacation periods, alleviating congestion. Likewise, allowing more customers to renege without service reduces the queue length and wait times.
- The positive effect of operator breakdown rate (φ) on expected cost is expected, since more breakdowns directly degrade service capability and capacity. In contrast and as anticipated, faster operator repair rate (γ) significantly improves system performance and reduces costs by restoring capacity quicker after failures.
- Larger system capacity (L) and impatience rates ($\varsigma_j, j = 0, 1, 2$) increase service abandonment, lowering congestion and $\mathcal{T}_c(M^*, c^*, \mu^*, \nu^*)$. However, excessive abandonment negatively impact customer service. An optimal balance is required.

3.7 Conclusion

In this chapter, based on the characteristics of the repair machine, we presented a $M/M/c/L$ queue with breakdowns, repairs, threshold-based recovery policy, working vacation, Bernoulli interruption, balking, reneging, and retention. We established the steady-state solution of the system using Q-matrix. Then, we studied important system characteristics based on the steady-state probabilities. Finally, we presented the sensitivity and cost optimization analysis; we discussed an economic analysis as well as the optimal threshold, the optimal number of servers as well as the service rates μ and ν under a given cost assumption because determining these parameters to achieve the minimum cost is very important in queueing theory. As further potential future study, we can generalize this queueing model with to some different cases, as follows:

- (i) Considering the feedback phenomenon within the queueing systems, it is pertinent to examine the scenario involving feedback customers in the proposed queueing model.

- (ii) It will be interesting to incorporate retrial policy and preemptive resume priority, this makes the system closer to real-life congestion scenarios and the study can provide potentially practical application in flexible manufacturing systems, transportation system, telecommunication systems, and so on.

- (iii) One could also extend the present study by considering multi-optional services.

Chapter 4

Finite-capacity multi-server queues with catastrophes, repairs, Bernoulli feedback, and fluid approximations

4.1 Introduction

Accurately modeling disruptions and their impact on operations is crucial for effective planning and decision-making in various industries. This paper investigates a finite-capacity multi-server queueing system with Bernoulli feedback, incorporating catastrophes and repairs to reflect the unpredictable nature of disruptions in real-world operational environments. By employing a combination of techniques, including transient and steady-state analyses, as well as fluid queue approximations, we provide quantitative characterizations of the system's behavior under different scenarios. The proposed model and analytical methods offer practical solutions for operational decision-makers, enabling proactive risk management and improved system performance.

Queueing models are mathematical tools that can be employed to represent and analyze various queueing situations that arise in different domains including supply chain management, hospitals, computer and communication networks, and manufacturing and production processes. One of the features that some queueing models incorporate is catastrophes, which are events that cause the complete destruction of all units present in a queueing system. Queueing models with catastrophes are relevant for many practical applications where such events can have significant consequences. Several studies have explored the properties and applications of queueing models with catastrophes in different contexts, such as computer networks [37], broadband communication networks [66], muscle contraction processes [44], insect populations [65], emergency supply chain systems [61], and emergency systems with uncertainty [67], and communication networks and cellular networks [105].

Feedback in queueing systems is typically facilitated through repeated service, and evaluating the performance of such systems involves assessing the rate at which customers return for service. Customers may choose to seek re-service for diverse reasons, such as incomplete or unsatisfactory primary service. This phenomenon is commonly observed in communication networks, where distorted data may require repeated transmission to reach the intended recipient. Repeated service in queueing systems can be classified into two types: instant and delayed feedback. In the case of instant feedback, customers either immediately seek re-service or abandon the system following the completion of their initial service, according to a Bernoulli scheme. On the other hand, in the case of delayed feedback, customers also follow a Bernoulli scheme but may choose to enter an orbit, where they wait for a certain positive (random) duration before returning for re-service or ultimately abandoning the system. Such feedback schemes are referred to as Bernoulli-type feedback. The pioneering work by [134, 133] represented the initial studies focused on the analysis of both types of feedback systems. These studies utilized the generating function method to analyze two-dimensional Markov models of single-server queueing systems with an unlimited queue and an infinite orbit volume in the case of delayed feedback. In recent years, various authors have extensively studied these models, often analyzing systems with and without an orbit separately. For instance, models with instant feedback have been examined in [27, 96, 23, 29], while models with delayed feedback have been investigated in [109, 9, 20]. Concurrently, studies such as [75, 94, 100] have explored queueing system models with both types of feedback.

Analyzing queueing systems over an extended time period reveals that their characteristics become independent of initial conditions, resulting in a stationary state known as steady-state. Steady-state solutions represent time-independent solutions to queuing models. They are useful for designing and evaluating queueing systems that operate under stable conditions. However, steady-state solutions may not be applicable or realistic for many practical situations where the state of the system changes over time due to various factors, such as opening and closing times, seasonal variations, or random events. In such situations, transient solutions are more appropriate and informative. These latter are time-dependent and consider the influence of initial conditions on the system's behavior. They can provide the probabilities of different states of the system at any given time and capture the dynamics and evolution of the system over time. Although the early focus of queueing theory was on steady-state behavior due to the design of many queueing systems, recent years have seen significant progress in understanding transient behavior. This is particularly relevant in modern applications such as broadband networks [99], adaptive routing, and feedback-based rate control mechanisms [17].

Transient solutions are more realistic and useful for many practical applications where the system changes over time, such as broadband networks Nagarajan and Kurose[99], adaptive routing, and feedback-based rate control mechanisms Bonomi et al.[17], as well as business operations or service providers that open and close. Therefore, studying the transient behavior of queueing systems is important both from a theoretical and a practical perspective. Several researchers analyzed the transient solutions of different queueing models under various scenarios. For instance, Ammar[4] and Ammar[5] explored the transient solutions of single server multiple vacation queues with impatient customers and waiting server, respectively. Then, Sudhesh et al.[129] and Sudhesh and Azhagappan[127] investigated single server Markovian working vacation queues with heterogeneous service and variant impatience, respectively. Later, Suranga et al.[132] presented the transient solution of $M/M/1$ queue with differentiated vacations and impatient customers. Recently, Vadivukarasi and Kalidass[144] discussed single server Markovian multiple variant vacation queues.

A stochastic fluid flow system is a sort of input-output model that uses a continuous fluid to represent the input, which flows into and out of a buffer storage device at rates that vary randomly. This model is particularly useful in situations where the arrival process is made up of discrete units, but the inter-arrival time between successive units is negligible. In such cases, the flow of units can be approximated by a continuous fluid, since the individual units have a relatively small impact on the system's performance. The fluid buffer in these models can be filled, depleted, or both, depending on the current state of the underlying queueing model, which determines the rates at which fluid enters and leaves the buffer. Markov modulated fluid queues are a specific type of fluid models that find practical applications in the modeling of various physical phenomena. One of the advantages of such models is that they are often amenable to tractable analysis. Several real-world scenarios where Markov Modulated Fluid Flow models have proven valuable can be found in literature references [142, 76, 152]. The analysis of fluid queueing models operating in a stochastic environment has been examined by numerous researchers, with a particular focus on stationary analysis, which has been well-established [147, 18, 146]. The study of fluid queueing systems has gained significant attention from researchers in recent years due to the broad range of applications in various fields such as in computer and communication systems [15, 82], manufacturing systems [97], congestion control [145], risk processes [114], and so on.

Given the practical applications of catastrophes, we focus, in this chapter, on studying a $M/M/c/N$ queueing model with Bernoulli feedback, catastrophes and repairs. We derive the transient and steady-state solutions for this model and obtain the expression for the time-dependent and time-independent expected number of customers in the system. We also model the system as a fluid queue and analyze its stationary solution. The current research is motivated by several factors, which can be summarized as follows:

- Firstly, multi-server queuing models present additional complexity compared to single-server counterparts due to simultaneous operation of multiple servers, leading to more complex interactions between customers and servers. Consequently, analyzing and modeling multi-server queuing models necessitates meticulous attention to ensure accurate results, making them more challenging to study than single-server queueing models. However, despite the increased complexity, multi-server queuing models are crucial for modeling real-world systems such as call centers, hospitals, and manufacturing facilities, where multiple servers handle substantial customer volumes or tasks.

- Secondly, incorporating Bernoulli feedback in the system affords us the opportunity to comprehensively capture the intricate dynamics of recurring customers in the system. By modelling the feedback mechanism as a Bernoulli random variable, we are able to effectively account for the probability of customers returning to the system and how this affects the system performance.

- Additionally, the study of transient and steady state solutions in a multi-server queueing model with disaster and repairs holds substantial relevance in diverse real-world applications. For instance, within manufacturing systems, machine breakdowns requiring repairs can lead to disruptions in production. Likewise, in service systems such as call centers, the occurrence of virus attacks can significantly impact the overall efficiency of customer service due to downtime.

- Moreover, understanding the behavior of such systems enables the design of their performance. Analyzing the transient behavior provides valuable perceptions into system dynamics during the initial startup phase, when there may be a backlog of customers awaiting service. Meanwhile, studying the steady-state behavior offers insights into system performance over extended time periods, when it has achieved a stable operational state.

- Furthermore, investigating the fluid queue driven by the proposed model improves our understanding of system behavior. By examining the fluid queue, we can gain valuable perspectives into the system's average behavior over time, facilitating the design of the system for improved performance.

The chapter is structured as follows: The mathematical description of the system is presented in Section 2. Section 3 is devoted the transient-analysis of the queueing system. In Section 4, we present the steady-sate solution while in Section 5 we deal with the steady-state analysis of the fluid queue.

4.2 Mathematical description of the system

Consider a finite capacity multi-server queueing system modeled by the process $\{X(t), t \geq 0\}$, where $X(t)$ represents the number of customers in the system at time t. Customers arrive according to a Poisson process with a rate of τ, and service times follow an exponential distribution with parameter μ. In the case of unsatisfactory service, customers may provide feedback and re-enter the system with a probability of θ' or exit the system with a complementary probability of θ. The system has a maximum capacity of N customers, and the service discipline follows a First-Come-First-Served (FCFS) approach. Additionally, the occurrence of catastrophes is assumed to follow a Poisson process with a rate of ϕ. When a catastrophe occurs, all customers, including the one currently being served, are lost, and the system transitions to a repair state. The repair time for a failed server is exponentially distributed with a parameter of η. During the repair state, customers who arrive are not permitted to join the queue. The state transition diagram of the modulating process is depicted in Figure 4.1.

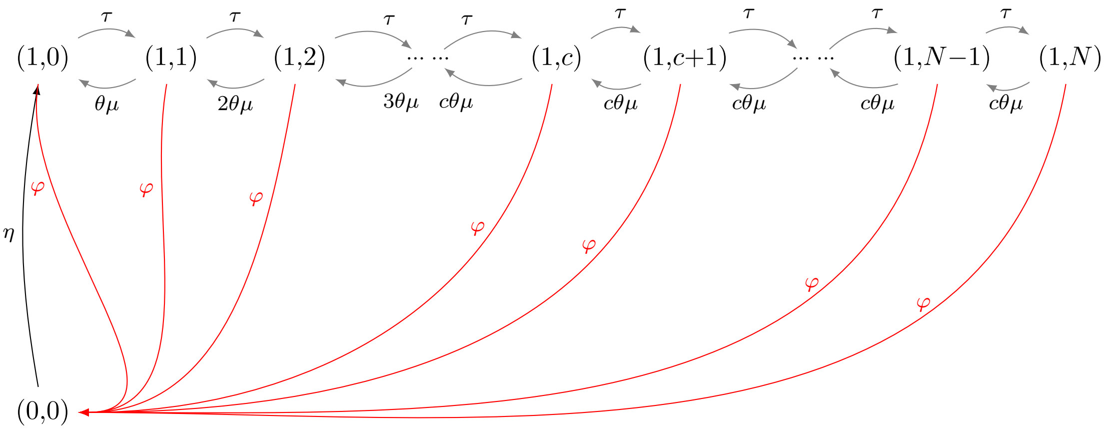

Figure 4.1: Transition diagram

4.3 Transient analysis

Let $X(t)$ denote the number of customers in the system at time t. Let $P_n(t) = \mathbb{P}(X(t) = n)$, $n = 0, 1, 2, ..., N$ represent the transient state probability that there are n customers in the system at time t when the server is operational. Let $R(t)$ be the probability that the server is under repair at time t.

Let $P(z,t) = \sum_{n=0}^{N} P_n(t) z^n$. Then, the differential-difference equations of the suggested queueing model are given as:

$$\begin{aligned}
\frac{dR(t)}{dt} &= -\eta R(t) + \varphi \sum_{n=0}^{N} P_n(t), && (4.1)\\
\frac{dP_0(t)}{dt} &= -(\tau+\varphi)P_0(t) + \theta\mu P_1(t) + \eta R(t), && (4.2)\\
\frac{dP_n(t)}{dt} &= -(\tau+\varphi+n\theta\mu)P_n(t) + \tau P_{n-1}(t) + (n+1)\theta\mu P_{n+1}(t), && (4.3)\\
& \quad 1 \le n \le c-1, \\
\frac{dP_c(t)}{dt} &= -(\tau+\varphi+c\theta\mu)P_c(t) + \tau P_{1n-1}(t) + c\theta\mu P_{c+1}(t), && (4.4)\\
\frac{dP_n(t)}{dt} &= -(\tau+\varphi+c\theta\mu)P_n(t) + \tau P_{c-1}(t) + c\theta\mu P_{n+1}(t), && (4.5)\\
& \quad c \le n \le N-1, \\
\frac{dP_N(t)}{dt} &= -(\varphi+c\theta\mu)P_N(t) + \tau P_{N-1}(t). && (4.6)
\end{aligned}$$

The normalizing condition is as:

$$R(t) + \sum_{n=0}^{N} P_n(t) = 1.$$

4.3.1 Evaluating $P_n(t)$,n=0,...,N and $R(t)$

The system is presumed to have no customers at the beginning, that is $P_0(0) = 1,\ P_n(t) = 0, n = 1, 2, ..., N$, and $R(0) = 0$. By solving Eq. 4.1 for $R(t)$ using the initial condition $R(0) = 0$, we obtain

$$R(t) = \frac{\varphi}{\eta+\varphi}\left[1 - e^{-(\eta+\varphi)t}\right]. \qquad (4.7)$$

The differential equation satisfied by the probability generating function (p.g.f) $P(z,t)$ can be readily observed as:

$$\begin{aligned}
\frac{\partial P(z,t)}{\partial t} &= \left[\tau z + \frac{c\theta\mu}{z} - (\tau+\varphi+c\theta\mu)\right] P(z,t) + c\theta\mu\left[1 - \frac{1}{z}\right] P_0(t)\\
&\quad + z^N \tau(1-z)P_N(t) + \eta R(t) + \theta\mu(1-z)\psi(z), \qquad (4.8)
\end{aligned}$$

where

$$\psi(z) = \sum_{n=1}^{c-1}(n-c)P_n(t)z^{n-1},$$

with the initial conditions $P(z,0)=1\ \ R(0)=0$. The solution of this partial differential equation is obtained as:

$$
\begin{aligned}
P(z,t) \;=\;& \eta\int_0^t R(t-u)e^{\left[\tau z+\frac{c\theta\mu}{z}-(\tau+\varphi+c\theta\mu)\right]u}du\\
&+c\theta\mu\left[1-\frac{1}{z}\right]\int_0^t P_0(t-u)e^{\left[\tau z+\frac{c\theta\mu}{z}-(\tau+\varphi+c\theta\mu)\right]u}du\\
&+z^N\tau(1-z)\int_0^t P_N(t-u)e^{\left[\tau z+\frac{c\theta\mu}{z}-(\tau+\varphi+c\theta\mu)\right]u}du\\
&+\theta\mu(1-z)\int_0^t \psi(z)e^{\left[\tau z+\frac{c\theta\mu}{z}-(\tau+\varphi+c\theta\mu)\right]u}du\\
&+e^{\left[\tau z+\frac{c\theta\mu}{z}-(\tau+\varphi+c\theta\mu)\right]t}. \qquad (4.9)
\end{aligned}
$$

4.3.1.1 Bessel function

Applying the Bessel function identity, when $\alpha=2\sqrt{\tau c\theta\mu}$ and $\delta=\sqrt{\dfrac{\tau}{c\theta\mu}}$, we have

$$\exp\left\{\left(\tau z+\frac{c\theta\mu}{z}\right)t\right\}=\sum_{n=-\infty}^{\infty}(\delta z)^n I_n(\alpha u).$$

Here, $I_n(.)$ represents the modified Bessel function of the first kind. By substituting the given expression into Eq. (4.9) and comparing the coefficients of z^n on both sides, we obtain the following equation for $n=0,1,...,N$:

$$
\begin{aligned}
P_n(t) \;=\;& \eta\delta^n\int_0^t R(t-u)e^{-(\tau+\varphi+c\theta\mu)u}I_n(\alpha u)du\\
&+c\theta\mu\delta^n\int_0^t P_0(t-u)e^{-(\tau+\varphi+c\theta\mu)u}\left[I_n(\alpha u)-\delta I_{n+1}(\alpha u)\right]du\\
&+\tau\delta^n\int_0^t P_N(t-u)e^{-(\tau+\varphi+c\theta\mu)u}\left[\delta^{-N}I_{N-n}(\alpha u)-\delta^{-(N+1)}I_{N+1-n}(\alpha u)\right]du\\
&+\theta\mu\delta^n\int_0^t\sum_{m=1}^{c-1}(m-c)P_m(t-u)e^{-(\tau+\varphi+c\theta\mu)u}\\
&\left[\delta^{-(m-1)}I_{m-n-1}(\alpha u)-\delta^{-m}I_{m-n}(\alpha u)\right]du+\delta^n I_n(\alpha t)e^{-(\tau+\varphi+c\theta\mu)t}. \qquad (4.10)
\end{aligned}
$$

In the previous step, we utilized the property that the modified Bessel function of the first kind satisfies the symmetry property:

$$I_{-n}(\cdot)=I_n(\cdot).$$

This allows us to simplify the equations derived from Equation 4.9 by considering only the cases where n is non-negative.

4.3.1.2 Laplace transform.

We obtained expressions for $P_n(t)$, where n ranges from 1 to $N-1$. However, these expressions are dependent on $P_0(t)$, $P_N(t)$, and $R(t)$. Fortunately, Eq. 4.7 provides an explicit expression for $R(t)$. To determine the values of $P_0(t)$ and $P_N(t)$, we utilize the Laplace transform. In the following discussion, for any function $f(\cdot)$, the Laplace transform of f is denoted as $f^*(s)$, given by:

$$f^*(s) = \int_0^\infty e^{-st} f(t)dt.$$

Here, the Laplace transform enables us to transform functions of time t into functions of the complex variable s. Substituting $n = 0$ into Eq. 4.10 we have:

$$\begin{aligned}
P_0(t) &= \eta \int_0^t R(t-u)e^{-(\tau+\varphi+c\theta\mu)u} I_0(\alpha u)du \\
&+c\theta\mu \int_0^t P_0(t-u)e^{-(\tau+\varphi+c\theta\mu)u} \left[I_0(\alpha u) - \delta I_1(\alpha u)\right] du \\
&+\tau \int_0^t P_N(t-u)e^{-(\tau+\varphi+c\theta\mu)u} \left[\delta^{-N} I_N(\alpha u) - \delta^{-(N+1)} I_{N+1}(\alpha u)\right] du \\
&+\theta\mu \int_0^t \sum_{m=1}^{c-1} (m-c)P_m(t-u)e^{-(\tau+\varphi+c\theta\mu)u} \\
&\left[\delta^{-(m-1)} I_{m-1}(\alpha u) - \delta^{-m} I_m(\alpha u)\right] du \\
&+I_0(\alpha t)e^{-(\tau+\varphi+c\theta\mu)t}. \qquad (4.11)
\end{aligned}$$

By taking the Laplace transform on both sides of Eq. (4.11) and solving for $P_0(s)$, we get

$$\begin{aligned}
\sqrt{\kappa^2-\alpha^2} P_0^*(s) &= \eta R^*(s) + c\theta\mu \left[1 - \frac{\kappa - \sqrt{\kappa^2-\alpha^2}}{2c\theta\mu}\right] P_0^*(s) \\
&+\tau \left[\frac{\kappa - \sqrt{\kappa^2-\alpha^2}}{2\tau}\right]^N \left[1 - \frac{\kappa - \sqrt{\kappa^2-\alpha^2}}{2\tau}\right] P_N^*(s) \\
&+\theta\mu \sum_{m=1}^{c-1} (m-c)P_m^*(s) \\
&\left[\frac{\kappa - \sqrt{\kappa^2-\alpha^2}}{2\tau}\right]^m \left[\frac{2\tau}{\kappa - \sqrt{\kappa^2-\alpha^2}} - 1\right] + 1, \quad (4.12)
\end{aligned}$$

where $\kappa = s + \tau + \varphi + c\theta\mu$. After some algebraic manipulation, the equation can be expressed as:

$$
\begin{aligned}
\left\{\sqrt{\kappa^2-\alpha^2}-c\theta\mu\left[1-\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2c\theta\mu}\right]\right\}P_0^*(s) \quad = \quad & \tau\left[\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2\tau}\right]^N \\
& \left[1-\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2\tau}\right]P_N^*(s) \\
& +\theta\mu\sum_{m=1}^{c-1}(m-c)P_m^*(s)\left[\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2\tau}\right]^m \\
& \left[\frac{2\tau}{\kappa-\sqrt{\kappa^2-\alpha^2}}-1\right]+1+\eta R^*(s). \quad (4.13)
\end{aligned}
$$

Then, by solving Eq. (4.13), we obtain:

$$
\begin{aligned}
P_0^*(s) \quad = \quad & \frac{\tau}{c\theta\mu}\left[\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2\tau}\right]^{N+1} P_N^*(s) \\
& +\frac{1}{c}\sum_{m=1}^{c-1}(m-c)P_m^*(s)\left[\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2\tau}\right]^m \\
& +\frac{1}{c\theta\mu}\left[\eta R^*(s)+1\right] \\
& \left[\frac{c\theta\mu}{s+\varphi}-\frac{\tau}{s+\varphi}\left(\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2\tau}\right)\right]. \qquad (4.14)
\end{aligned}
$$

On inversion, this equation yields an expression for $P_0(t)$ that depends on $P_0^*(t)$:

$$
\begin{aligned}
P_0(t) \quad = \quad & e^{-\varphi t}+\eta\int_0^t R(t-u)e^{-\varphi u}du-\eta\delta\left[e^{-\varphi t}*R(t)*e^{-(\tau+\varphi+c\theta\mu)t}\frac{I_1(\alpha t)}{t}\right] \\
& -\delta\int_0^t e^{-\varphi(t-u)}e^{-(\tau+\varphi+c\theta\mu)u}\frac{I_1(\alpha u)}{u} \\
& +\left(\frac{c\theta\mu}{\tau}\right)^{\frac{N-1}{2}}\int_0^t P_N(t-u)e^{-(\tau+\varphi+c\theta\mu)u}(N+1)\frac{I_{N+1}(\alpha u)}{u}du \\
& +\frac{1}{c}\int_0^t\sum_{m=1}^{c-1}(m-c)P_m(t-u)*e^{-(\tau+\varphi+c\theta\mu)u}m\delta^{-m}\frac{I_m(\alpha u)}{u}du, \quad (4.15)
\end{aligned}
$$

where $*$ denotes the convolution operation. Next, substituting $n = N$

into Eq. (4.10), we obtain

$$\begin{aligned} P_N(t) &= \eta\delta^N \int_0^t R(t-u)e^{-(\tau+\varphi+c\theta\mu)u} I_N(\alpha u)du \\ &+c\theta\mu\delta^N \int_0^t P_0(t-u)e^{-(\tau+\varphi+c\theta\mu)u}\left[I_N(\alpha u)-\delta I_{N+1}(\alpha u)\right]du \\ &+\tau \int_0^t P_N(t-u)e^{-(\tau+\varphi+c\theta\mu)u}\left[I_0(\alpha u)-\delta^{-1}I_1(\alpha u)\right]du \\ &+\theta\mu\delta^N \int_0^t \sum_{m=1}^{c-1}(m-c)P_m(t-u)e^{-(\tau+\varphi+c\theta\mu)u} \\ &\left[\delta^{-(m-1)}I_{m-N-1}(\alpha u)-\delta^{-m}I_{m-N}(\alpha u)\right]du \\ &+\delta^N I_N(\alpha t)e^{-(\tau+\varphi+c\theta\mu)t}. \end{aligned} \tag{4.16}$$

By taking Laplace transforms and solving for $P_N(s)$, we obtain the following expression from equation (4.16):

$$\begin{aligned} \sqrt{\kappa^2-\alpha^2}P_N^*(s) &= \eta R^*(s)\left[\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2c\theta\mu}\right]^N \\ &+c\theta\mu\left[\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2c\theta\mu}\right]^N\left[1-\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2c\theta\mu}\right]P_0^*(s) \\ &+\theta\mu\sum_{m=1}^{c-1}(m-c)P_m^*(s)\left[\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2\tau}\right]^{m-N} \\ &\left[\frac{2\tau}{\kappa-\sqrt{\kappa^2-\alpha^2}}-1\right]+\left[\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2c\theta\mu}\right]^N \\ &+\tau\left[1-\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2\tau}\right]P_N^*(s). \end{aligned} \tag{4.17}$$

Then, by substituting Eq. (4.14) into (4.17), we have:

$$\begin{aligned} &\left\{\sqrt{\kappa^2-\alpha^2}-\tau+\left[\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2}\right]\right\}P_N^*(s)-\tau\left[\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2c\theta\mu}\right]^N\left[\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2\tau}\right]^{N+1} \\ &\left[1-\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2c\theta\mu}\right]P_N^*(s)=\left[1+\left[1-\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2c\theta\mu}\right]\left[\frac{c\theta\mu}{s+\varphi}-\frac{\tau}{s+\varphi}\left(\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2\tau}\right)\right]\right] \\ &\times(\eta R^*(s)+1)\left[\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2c\theta\mu}\right]^N+\theta\mu\sum_{m=1}^{c-1}(m-c)P_m^*(s)\left[\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2\tau}\right]^m \\ &\times\left\{\left[1-\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2c\theta\mu}\right]\left[\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2c\theta\mu}\right]^N+\left[\frac{2\tau}{\kappa-\sqrt{\kappa^2-\alpha^2}}\right]^N\left[\frac{2\tau}{\kappa-\sqrt{\kappa^2-\alpha^2}}-1\right]\right\} \end{aligned} \tag{4.18}$$

After simplification, Eq. (4.18) can be done as:

$$(1-f^*(s))P_N^*(s)=\mathcal{G}^*(s)+\mathcal{K}^*(s), \tag{4.19}$$

where

$$\begin{aligned} f^*(s) &= \left[\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2c\theta\mu}\right]^{N+1}\left[\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2\tau}\right]^{N+1}\left[1-\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2c\theta\mu}\right] \\ &+\left[\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2c\theta\mu}\right], \end{aligned} \tag{4.20}$$

$$\begin{aligned} \mathcal{G}^*(s) &= \frac{1}{\tau}(\eta R^*(s)+1)h^*(s) \\ &= \frac{1}{\tau}(\eta R^*(s)+1)\left[\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2c\theta\mu}\right]^{N+1} \\ &\quad \left[1+\left[1-\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2c\theta\mu}\right]\right. \\ &\quad \left.\left[\frac{c\theta\mu}{s+\varphi}-\frac{\tau}{s+\varphi}\left(\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2\tau}\right)\right]\right], \end{aligned} \tag{4.21}$$

with

$$h^*(s)=\left[\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2c\theta\mu}\right]^{N+1}\left[1+\left[1-\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2c\theta\mu}\right]\left[\frac{c\theta\mu}{s+\varphi}-\frac{\tau}{s+\varphi}\left(\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2\tau}\right)\right]\right],$$

and

$$\begin{aligned} \mathcal{K}^*(s) &= \frac{\theta\mu}{\tau}\sum_{m=1}^{c-1}(m-c)P_m^*(s)\left[\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2\tau}\right]^m \\ &\quad \left[\left[1-\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2c\theta\mu}\right]\left[\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2c\theta\mu}\right]^{N+1}\right. \\ &\quad +\left[\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2c\theta\mu}\right]^{N+1}\left[\frac{2\tau}{\kappa-\sqrt{\kappa^2-\alpha^2}}\right]^N \\ &\quad \left.\left[\frac{2\tau}{\kappa-\sqrt{\kappa^2-\alpha^2}}-1\right]\right]. \end{aligned} \tag{4.22}$$

Upon inverting equations (4.20) and (4.21)-(4.22) we obtain an expression for $f(t)$, $h(t)$, $\mathcal{G}(t)$, and $\mathcal{K}(t)$ as follows:

$$\begin{aligned} f(t) &= \sqrt{\frac{\tau}{c\theta\mu}}e^{-(\tau+\varphi+c\theta\mu)t}\frac{I_1(\alpha t)}{t}+e^{-(\tau+\varphi+c\theta\mu)t}(2N+2)\frac{I_{2N+2}(\alpha t)}{t} \\ &-\sqrt{\frac{\tau}{c\theta\mu}}e^{-(\tau+\varphi+c\theta\mu)t}(2N+3)\frac{I_{2N+3}(\alpha t)}{t}, \end{aligned} \tag{4.23}$$

$$
\begin{aligned}
h(t) &= \left(\frac{\tau}{c\theta\mu}\right)^{\dfrac{N+1}{2}} e^{-(\tau+\varphi+c\theta\mu)t}(N+1)\frac{I_{N+1}(\alpha t)}{t} \\
&+ \left(\frac{\tau}{c\theta\mu}\right)^{\dfrac{N+1}{2}} * \int_0^t e^{-\varphi(t-u)}e^{-(\tau+\varphi+c\theta\mu)u} \\
&\left[c\theta\mu(N+1)\frac{I_{N+1}(\alpha u)}{u} - 2\sqrt{\tau c\theta\mu}(N+2)\frac{I_{N+2}(\alpha u)}{u}\right] du \\
&+ \left(\frac{\tau}{c\theta\mu}\right)^{\dfrac{N+1}{2}} \tau \int_0^t e^{-\varphi(t-u)}e^{-(\tau+\varphi+c\theta\mu)u}(N+3) \\
&\frac{I_{N+3}(\alpha u)}{u} du, \qquad (4.24)
\end{aligned}
$$

$$
\mathcal{G}(t) = \frac{\eta}{\tau}\int_0^t R(t-u)h(u)du + \frac{1}{\tau}h(t), \qquad (4.25)
$$

and

$$
\begin{aligned}
\mathcal{K}(t) &= \frac{\theta\mu}{\tau}\int_0^t \sum_{m=1}^{c-1}(m-c)P_m(u)e^{-(\tau+\varphi+c\theta\mu)u} \\
&\times \Bigg[\delta^{N+1-m}(N+m+1)\frac{I_{N+m+1}(\alpha u)}{u} - \\
&\delta^{N-m+2}(N+m+2)\frac{I_{N+m+2}(\alpha u)}{u} \\
&+\delta^{N+2-m}(m-N)\frac{I_{m-N}(\alpha u)}{u} - \\
&\delta^{N+1-m}(m-N+1)\frac{I_{m-N+1}(\alpha u)}{u}\Bigg]. \qquad (4.26)
\end{aligned}
$$

Given that $0 \leqslant f^*(s) < 1$, we can express Eq. (4.19) as follows:

$$
P_N^*(s) = \mathcal{H}^*(s)\sum_{i=0}^{\infty}\left[f^*(s)\right]^i,
$$

where $\mathcal{H}^*(s) = \mathcal{G}^*(s) + K^*(s)$. Upon inversion, this equation provides an expression for $P_N^*(t)$ as follows:

$$
P_N(t) = H(t) * \sum_{i=0}^{\infty}\left[f(t)\right]^{*i}. \qquad (4.27)
$$

Here, $[f(t)]^{*i}$ represents the i-fold convolution of $f(t)$ with it self. Note that $[f(t)]^{*0} = 1$.

4.3.2 Time dependent mean and variance of the system size

Lemme 4.3.1. *The time dependent expression for the mean of the system size is given by:*

$$\begin{aligned}\mathcal{E}(t) &= \frac{1}{\varphi}(\tau - c\theta\mu)(1-e^{-\varphi t}) - \tau\int_0^t P_N(u)e^{-\varphi(t-u)}du \\ &+(c\theta\mu-\tau)\int_0^t R(u)e^{-\varphi(t-u)}du \\ &+c\theta\mu\left(\int_0^t P_0(u)e^{-\varphi(t-u)}du - \int_0^t P_c(u)e^{-\varphi(t-u)}du\right) \\ &+\int_0^t \left[\phi_2(u)-\phi_1(u)\right]e^{-\varphi(t-u)}du. \end{aligned} \tag{4.28}$$

Proof. Note that $\mathcal{E}(t) = \mathbb{E}(X(t)) = \sum_{n=1}^{N} nP_n(t)$, $\mathcal{E}(0) = \sum_{n=1}^{N} nP_n(0) = 0$, and $\mathcal{E}'(t) = \sum_{n=1}^{N} nP_n'(t)$. Then, making use of Eqs. (4.4) and (4.6), we have

$$\begin{aligned}\mathcal{E}'(t) &= -\tau\sum_{n=1}^{N-1} nP_n(t) - \varphi\sum_{n=1}^{N} nP_n(t) - n\theta\mu\sum_{n=1}^{c-1} nP_n(t) - c\theta\mu\sum_{n=c}^{N} nP_n(t) \\ &+\tau\sum_{n=1}^{N} nP_{n-1}(t) + (n+1)\theta\mu\sum_{n=1}^{c-1} nP_{n+1}(t) + c\theta\mu\sum_{n=c}^{N-1} nP_{n+1}(t).\end{aligned}$$

After performing some algebraic manipulations, the aforementioned equation can be expressed as follows:

$$\begin{aligned}\mathcal{E}'(t) &= -\varphi\mathcal{E}(t) - \tau P_N(t) + (\tau - c\theta\mu) + (c\theta\mu-\tau)R(t) + c\theta\mu P_0(t) - c\theta\mu P_c(t) \\ &\quad -\phi_1(t)+\phi_2(t),\end{aligned} \tag{4.29}$$

where $\phi_1(t) = \theta\mu\sum_{n=1}^{c-1}(n-c)P_n(t)$ and $\phi_2(t) = \theta\mu\sum_{n=1}^{c-1} n^2P_n(t)$. Considering Equation (4.29) as a first-order linear differential equation in $\mathcal{E}(t)$, we can find the solution by identifying the integrating factor and utilizing the initial condition $\mathcal{E}(0) = 0$. Consequently, the solution is obtained as Equation (4.28). □

Lemme 4.3.2. *The time dependent expression for the variance of the system* *size is given by:*

$$
\begin{aligned}
Var(X(t)) &= \frac{1}{\varphi}(\tau + c\theta\mu)(1 - e^{-\varphi t}) - \tau(2N+1)\int_0^t P_N(u)e^{-\varphi(t-u)}du \\
&\quad -(\tau + c\theta\mu)\int_0^t R(u)e^{-\varphi(t-u)}du \\
&\quad +2(\tau - c\theta\mu)\int_0^t M(u)e^{-\varphi(t-u)}du \\
&\quad -c\theta\mu\int_0^t P_0(u)e^{-\varphi(t-u)}du - c^3\theta\mu\int_0^t P_c(u)e^{-\varphi(t-u)}du \\
&\quad +2(c-1)\int_0^t \phi_2(u)e^{-\varphi(t-u)}du \\
&\quad +\int_0^t \left[\phi_1(u) + \phi_3(u)\right] e^{-\varphi(t-u)}du - [M(t)]^2. \qquad (4.30)
\end{aligned}
$$

Proof. It is known that

$$
\begin{aligned}
Var(X(t)) &= \mathbb{E}(X^2(t)) - [\mathbb{E}(X(t))]^2 \\
&= \mathcal{V}(t) - [\mathcal{E}(t)]^2, \qquad (4.31)
\end{aligned}
$$

where

$$
\mathcal{V}(t) = \mathbb{E}(X^2(t)) = \sum_{n=1}^{N} n^2 P_n(t),\ \ \mathcal{V}(0) = \sum_{n=1}^{N} n^2 P_n(0) = 0,\ \text{ and }\ \mathcal{V}'(t) = \sum_{n=1}^{N} n^2 P_n'(t).
$$

From Eqs. (4.4) and (4.6) we have

$$
\begin{aligned}
\mathcal{V}'(t) &= -\tau\sum_{n=1}^{N-1} n^2 P_n(t) - \varphi\sum_{n=1}^{N} n^2 P_n(t) - n\theta\mu\sum_{n=1}^{c-1} n^2 P_n(t) - c\theta\mu\sum_{n=c}^{N} n^2 P_n(t) \\
&\quad +\tau\sum_{n=1}^{N} n^2 P_{n-1}(t) + (n+1)\theta\mu\sum_{n=1}^{c-1} n^2 P_{n+1}(t) + c\theta\mu\sum_{n=c}^{N-1} n^2 P_{n+1}(t).
\end{aligned}
$$

After performing some algebraic manipulations, the above equation can be expressed as follows:

$$
\begin{aligned}
\mathcal{V}'(t) &= -\varphi\mathcal{V}(t) + (\tau + c\theta\mu) - (\tau + c\theta\mu)R(t) - \tau(2N+1)P_N(t) \\
&\quad +2(\tau - c\theta\mu)M(t) - c\theta\mu P_0(t) - c^3\theta\mu P_c(t) + \phi_1(t) \\
&\quad +2(c-1)\phi_2(t) + \phi_3(t), \qquad (4.32)
\end{aligned}
$$

where

$$
\phi_3(t) = 2c\theta\mu\sum_{n=1}^{c} nP_n(t).
$$

Considering the above equation as a first-order differential equation in $\mathcal{V}(t)$, we can find the solution by identifying the integrating factor and utilizing the initial condition $\mathcal{V}(0) = 0$. The solution of the equation is obtained as follows:

$$\begin{aligned}
\mathcal{V}(t) &= \frac{1}{\varphi}(\tau + c\theta\mu)(1 - e^{-\varphi t}) - \tau(2N+1)\int_0^t P_N(u)e^{-\varphi(t-u)}du \\
&\quad -(\tau + c\theta\mu)\int_0^t R(u)e^{-\varphi(t-u)}du \\
&\quad +2(\tau - c\theta\mu)\int_0^t M(u)e^{-\varphi(t-u)}du \\
&\quad -c\theta\mu\int_0^t P_0(u)e^{-\varphi(t-u)}du - c^3\theta\mu\int_0^t P_c(u)e^{-\varphi(t-u)}du \\
&\quad +2(c-1)\int_0^t \phi_2(u)e^{-\varphi(t-u)}du \\
&\quad +\int_0^t \left[\phi_1(u) + \phi_3(u)\right] e^{-\varphi(t-u)}du. \qquad (4.33)
\end{aligned}$$

Finally, substituting Eq. (4.33) into Eq. (4.31), we get (4.30). □

4.4 Steady-state analysis

This section focuses on a the discussion of the structure of steady state probabilities for the system size and failure distribution in a multi-server and finite capacity queueing system that incorporates factors such as Bernoulli feedback, disasters and non-zero repair times. Our analysis aims to examine the behavior of the system under these specific conditions. Additionally, we derive expressions for the steady state mean and variance.

Theorem 4.1. *For $\varphi > 0$ $\eta > 0$, we have:*

$$\begin{aligned}
Q &= \frac{\varphi}{\eta + \varphi}, \\
P_0 &= \rho\rho_1^N P_N + (1-Q)(1-\rho) + \frac{1}{c}\sum_{m=1}^{c-1}(m-c)\upsilon^m P_m, \\
P_n &= \sigma c\theta\mu\rho^{n+1}(1-\rho)\upsilon^N P_N + \sigma\tau\rho_1^{N-n}(1-\upsilon)P_N + (1-Q)\rho^n(1-\rho) \\
&\quad +\sigma\theta\mu\sum_{m=1}^{c-1}(m-c)\upsilon^m P_m\left[\rho^n(1-\rho) + \upsilon^{-(n+1)}(1-\upsilon)\right],
\end{aligned}$$

and

$$\begin{aligned} P_N &= \left(\rho^{N+1}(1-Q)(\varphi + c\theta\mu[1-\rho]^2) + \theta\mu \sum_{m=1}^{c-1} (m-c)\upsilon^m P_m \right. \\ &\quad \left. \times [\rho^{N+1}(1-\rho) + \rho\upsilon^{-N-1}(1-\upsilon)] \right) \\ &\quad \times \left\{ \tau(1-\rho-\rho_1^{N+1}\rho^{N+1}(1-\rho)) \right\}^{-1}, \end{aligned} \tag{4.34}$$

where

$$\rho = \frac{(\tau+\varphi+c\theta\mu) - \sqrt{(\tau+\varphi+c\theta\mu)^2 - 4\tau c\theta\mu}}{2c\theta\mu},$$

$$\upsilon = \frac{(\tau+\varphi+c\theta\mu) - \sqrt{(\tau+\varphi+c\theta\mu)^2 - 4\tau c\theta\mu}}{2\tau}, \tag{4.35}$$

and

$$\sigma = \frac{1}{\sqrt{(\tau+\varphi+c\theta\mu)^2 - 4\tau c\theta\mu}}, \tag{4.36}$$

Proof. For $\varphi > 0$ $\eta > 0$, taking limit at $t \longrightarrow \infty$ in Eq. (4.7), we get

$$Q = \frac{\varphi}{\eta + \varphi}. \tag{4.37}$$

Multiplying Eq. (4.14) by s on both sides and taking limit as $s \longrightarrow 0$, we have

$$\lim_{s \longrightarrow 0} sP_0^*(s) = \frac{\tau}{c\theta\mu}\rho_1^{N+1}P_N + \frac{\eta}{c\theta\mu}Q\left(\frac{c\theta\mu}{\varphi} - \frac{\tau}{\varphi}\upsilon\right) + \frac{1}{c}\sum_{m=1}^{c-1}(m-c)\upsilon^m P_m.$$

Then,

$$P_0 = \rho\rho_1^N P_N + \frac{\eta}{\varphi}R(1-\rho) + \frac{1}{c}\sum_{m=1}^{c-1}(m-c)\upsilon^m P_m.$$

After some algebra, we get

$$P_0 = \rho\rho_1^N P_N + (1-Q)(1-\rho) + \frac{1}{c}\sum_{m=1}^{c-1}(m-c)\upsilon^m P_m \tag{4.38}$$

By taking Laplace transforms of the Eq. (4.10), we get

$$\begin{aligned}P_n^*(s) &= c\theta\mu\frac{1}{\sqrt{\kappa^2-\alpha^2}}\left[\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2c\theta\mu}\right]^n\left[1-\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2c\theta\mu}\right]\\ &+\tau P_N^*(s)\frac{1}{\sqrt{\kappa^2-\alpha^2}}\left[\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2\tau}\right]^{N-n}\left[1-\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2\tau}\right]\\ &+\eta R^*(s)\frac{1}{\sqrt{\kappa^2-\alpha^2}}\left[\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2c\theta\mu}\right]^n+\frac{1}{\sqrt{\kappa^2-\alpha^2}}\left[\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2c\theta\mu}\right]^n\\ &+\frac{\theta\mu}{\sqrt{\kappa^2-\alpha^2}}\sum_{m=1}^{c-1}(m-c)P_m^*(s)\left[\frac{\kappa-\sqrt{\kappa^2-\alpha^2}}{2\tau}\right]^{m-n}\left[\frac{2\tau}{\kappa-\sqrt{\kappa^2-\alpha^2}}-1\right].\end{aligned}$$

Multiplying the above equation by s on both sides and taking limit as $s\longrightarrow 0$, we obtain

$$\begin{aligned}\lim_{s\longrightarrow 0} sP_n^*(s) &= \sigma c\theta\mu\rho^n(1-\rho)P_0+\sigma\tau\rho_1^{N-n}(1-\upsilon)P_N+\sigma\eta\rho^nQ\\ &\quad+\sigma\theta\mu\sum_{m=1}^{c-1}(m-c)\upsilon^{m-n-1}(1-\upsilon)P_m.\end{aligned}$$

Substituting Eq. (4.38) in the above equation, we get

$$\begin{aligned}P_n = \sigma c\theta\mu\rho^{n+1}(1-\rho)\upsilon^NP_N+\sigma\tau\rho_1^{N-n}(1-\upsilon)P_N+(1-Q)\rho^n(1-\rho)\\ +\sigma\theta\mu\sum_{m=1}^{c-1}(m-c)\upsilon^mP_m\left[\rho^n(1-\rho)+\upsilon^{-(n+1)}(1-\upsilon)\right]\end{aligned} \tag{4.39}$$

Multiplying Eq. (4.21) by s on both sides and taking limit as $s\longrightarrow 0$,

$$\lim_{s\longrightarrow 0} sG^*(s) = \frac{1}{\tau}(1-Q)(\varphi+c\theta\mu[1-\rho]^2)\rho^{N+1}. \tag{4.40}$$

Taking limit as $s\longrightarrow 0$ in Eq. (4.20), we have

$$\lim_{s\longrightarrow 0} f^*(s) = \rho(1+\rho_1^{N+1}\rho^N(1-\rho)). \tag{4.41}$$

Further, taking limit as $s\longrightarrow 0$ in Eq. (4.22), we find

$$\lim_{s\longrightarrow 0} sK^*(s) = \frac{\theta\mu}{\tau}\sum_{m=1}^{c-1}(m-c)\upsilon^mP_m\left[\rho^{N+1}(1-\rho)+\rho\upsilon^{-N-1}(1-\upsilon)\right]. \tag{4.42}$$

Multiplying Eqs. (4.19) by s on both sides and taking limit as $s\longrightarrow 0$, we obtain

$$\lim_{s\longrightarrow 0} sP_N^*(s) = \lim_{s\longrightarrow 0}\frac{sG^*(s)+sK^*(s)}{1-f^*(s)} \tag{4.43}$$

Substituting Eqs. (4.40) and (4.41) in the above equation, we get Eq. (4.34).

□

4.5 Analysis of fluid queue.

One practical application of considering this type of fluid queue is in the analysis of transportation systems, such as highways, airports, or seaports. These systems involve the movement of vehicles or passengers through a network of roads, runways, or waterways, where they may encounter congestion, delays, or accidents. By modeling the flow of vehicles or passengers as a fluid queue, we can capture the dynamics of the system and account for the occurrence of disasters that may disrupt the normal flow of traffic. Disasters can be natural events, such as storms, floods, or earthquakes, or human-caused events, such as terrorist attacks, strikes, or riots. These events can cause the loss of all vehicles or passengers in the system, either by destroying them or by forcing them to evacuate. By using a fluid queue model, we can estimate the probability and impact of such disasters on the system performance and reliability. For example, consider an airport with multiple runways and a finite capacity of parking spaces for planes. The airport can be modeled as a fluid queue operated by an $M/M/c/N$ queueing system with disasters, where the fluid represents the planes arriving and departing from the airport, and the service channels represent the number of runways available for takeoff and landing. The model can be used to analyze the impact of disasters, such as severe weather conditions or equipment failures, on the performance of the airport, including the waiting times, queue length, and resource utilization. We can also use the model to compare different scenarios and identify optimal solutions for improving the airport's capacity and resilience. For instance, we can evaluate how adding more runways or increasing the parking capacity would affect the average waiting time and queue length of planes. We can also assess how changing the disaster rate would affect the expected number of planes in the system at any given time. This information can be used to improve the overall effectiveness of transportation systems and ensure that resources are used efficiently during times of disruption.

Consider the buffer content process $\{\mathcal{C}(t);\ t \geqslant 0\}$, where $\mathcal{C}(t)$ is the amount of fluid in the buffer at time t. The fluid enters the buffer at a constant rate $\chi > 0$ when the servers are working (busy or idle), and leaves the buffer at a constant rate $\chi_0 < 0$ when the servers are under repair, as long as the buffer is not empty. The buffer content process has the following dynamics:

$$\frac{d\mathcal{C}(t)}{dt} = \begin{cases} 0, & \text{if } J(t) = 0,\ X(t) = 0,\ \mathcal{C}(t) = 0; \\ \\ \chi_0, & \text{if } J(t) = 0,\ X(t) = 0,\ \mathcal{C}(t) > 0; \\ \\ \chi, & \text{if } J(t) = 1,\ X(t) = i,\ i = 0, 1, 2, .., N,\ \mathcal{C}(t) \geqslant 0. \end{cases}$$

The process $\{(J(t), X(t), \mathcal{C}(t)), t \geqslant 0\}$ is a Markov process with three dimensions and has a unique stationary distribution under a suitable stability condition. We assume that the mean total input rate is negative, that is,

$$\chi_0 q + \chi \sum_{k=0}^{N} P_k < 0.$$

Define,

$$R(t,x) = P\left\{(J(t)=0, X(t)=0, \mathcal{C}(t) \leqslant x), t, x \geqslant 0\right\},$$

and

$$F_k(t,x) = P\left\{(J(t)=1, X(t)=k, \mathcal{C}(t) \leqslant x), t, x \geqslant 0, k=0,1,2,...,N\right\}.$$

The Markov process $\{J(t), X(t), \mathcal{C}(t)\}$ satisfies the Kolmogorov forward equations

$$\begin{aligned}
\frac{\partial R(t,x)}{\partial t} + \chi_0 \frac{\partial R(t,x)}{\partial x} &= -\eta R(t,x) + \varphi \sum_{n=0}^{N} F_k(t,x); \\
\frac{\partial F_0(t,x)}{\partial t} + \chi \frac{\partial F_0(t,x)}{\partial x} &= -(\tau+\varphi) F_0(t,x) + \theta\mu F_1(t,x) + \eta R(t,x); \\
\frac{\partial F_k(t,x)}{\partial t} + \chi \frac{\partial F_k(t,x)}{\partial x} &= -(\tau+\varphi+n\theta\mu) F_k(t,x) + \tau F_{k-1}(t,x) + (n+1)\theta\mu F_{k+1}(t,x); \\
&\quad k = 1,2,..,c-1. \\
\frac{\partial F_k(t,x)}{\partial t} + \chi \frac{\partial F_k(t,x)}{\partial x} &= -(\tau+\varphi+c\theta\mu) F_k(t,x) + \tau F_{k-1}(t,x) + c\theta\mu F_{k+1}(t,x); \\
&\quad k = c, c+1,..,N-1, \\
\frac{\partial F_N(t,x)}{\partial t} + \chi \frac{\partial F_N(t,x)}{\partial x} &= -(\varphi+c\theta\mu) F_N(t,x) + \tau F_{N-1}(t,x);
\end{aligned}$$

When the system reaches a steady state, such that $R(x)$ and $F_k(x)$ are the limits of $R(t,x)$ and $F_k(t,x)$ as t approaches infinity, respectively, for $k = 0,1,2,...,N$, the system simplifies to

$$\chi_0 \frac{dR(x)}{dx} = -\eta R(x) + \varphi \sum_{k=1}^{N} F_k(x), \tag{4.44}$$

$$\chi \frac{dF_0(x)}{dx} = -(\tau+\varphi) F_0(x) + \theta\mu F_1(x) + \eta R(x), \tag{4.45}$$

$$\chi \frac{dF_k(x)}{dx} = -(\tau+\varphi+n\theta\mu) F_k(x) + \tau F_{k-1}(x) + (n+1)\theta\mu F_{k+1}(x), \tag{4.46}$$

$$k = 1,2,..,c-1,$$

$$\chi \frac{dF_k(x)}{dx} = -(\tau+\varphi+c\theta\mu) F_k(x) + \tau F_{k-1}(x) + c\theta\mu F_{k+1}(x), \tag{4.47}$$

$$k = c, c+1,..,N-1, \tag{4.48}$$

$$\chi \frac{dF_N(x)}{dx} = -(\varphi+c\theta\mu) F_N(x) + \tau F_{N-1}(x),$$

subject to the boundary conditions, $F_k(0) = 0, k = 0,1,2,...,N$, and $R(0) = a$ for some constant $0 < a < 1$. The constant a such that $0 < a < 1$ is an unknown to be determined.

4.5.1 Stationary distribution

Assume that $R^*(s)$ and $F_k^*(s)$ are the Laplace transforms of $R(x)$ and $F_k(x)$, respectively.
Let the generating function be

$$G(z,x) = \sum_{k=0}^{N} F_k(x) z^k.$$

The system of difference-differential equations given by equations (4.45)-(4.49) results in a linear differential equation as follows:

$$\begin{aligned} \frac{\partial G(z,x)}{\partial t} &= \left[\frac{\tau z}{\chi} + \frac{c\theta\mu}{\chi z} - \frac{(\tau+\varphi+c\theta\mu)}{\chi}\right] G(z,x) \\ &+ \frac{c\theta\mu}{\chi}\left[1-\frac{1}{z}\right] F_0(x) + \frac{\tau}{\chi}(1-z)z^N F_N(x) \\ &+ \frac{\eta}{\chi} R(x) + \frac{\theta\mu}{\chi}(1-z)\psi"(z), \end{aligned} \tag{4.49}$$

where

$$\psi"(z) = \sum_{k=1}^{c-1} (k-c) F_k(x) z^{k-1}.$$

By integrating the above equation with respect to variable x, we get

$$\begin{aligned} G(z,x) &= \frac{\eta}{\chi}\int_0^t R(x-y) e^{\left[\frac{\tau z}{\chi} + \frac{c\theta\mu}{\chi z} - \frac{(\tau+\varphi+c\theta\mu)}{\chi}\right]y} dy \\ &+ \frac{c\theta\mu}{\chi}\left[1-\frac{1}{z}\right]\int_0^t F_0(x-y) e^{\left[\frac{\tau z}{\chi} + \frac{c\theta\mu}{\chi z} - \frac{(\tau+\varphi+c\theta\mu)}{\chi}\right]y} dy \\ &+ \frac{\tau}{\chi}(1-z)z^N \int_0^t F_N(x-y) e^{\left[\frac{\tau z}{\chi} + \frac{c\theta\mu}{\chi z} - \frac{(\tau+\varphi+c\theta\mu)}{\chi}\right]y} dy \\ &+ \frac{\theta\mu}{\chi}(1-z)\int_0^t \sum_{m=1}^{c-1} (m-c) F_m^*(x-y) z^{m-1} \\ & e^{\left[\frac{\tau z}{\chi} + \frac{c\theta\mu}{\chi z} - \frac{(\tau+\varphi+c\theta\mu)}{\chi}\right]y} dy. \end{aligned} \tag{4.50}$$

By setting $z = 1$ in Eq. (4.50), we obtain

$$\sum_{k=0}^{N} F_k(x) = \frac{\eta}{\chi}\int_0^x R(x-y) e^{-\frac{\varphi}{\chi}y} dy.$$

Applying Laplace transform to both sides of the equation, we get

$$\sum_{k=0}^{N} F_k^*(s) = \frac{\eta R^*(s)}{s\chi + \varphi}. \tag{4.51}$$

Likewise, Laplace transform of Eq. (4.44) gives

$$a\chi_0 + \sum_{k=0}^{N} F_k^*(s) = (s\chi_0 + \eta)R^*(s). \tag{4.52}$$

Using Eq. (4.51) to substitute into Eq. (4.52), and simplifying, we have

$$\begin{aligned} R^*(s) &= \frac{a\chi_0}{\chi_0 s + \eta - \dfrac{\eta\varphi}{\chi s + \varphi}} \\ &= a\chi_0 \sum_{j=0}^{\infty} \frac{\varphi^j \eta^j}{(\chi_0 s + \eta)^{j+1}(\chi s + \varphi)^j}, \end{aligned} \tag{4.53}$$

which can be inverted to give

$$R(x) = a \sum_{j=0}^{\infty} \left(\frac{\eta\varphi}{\chi_0 \chi}\right)^j \left[\left(\frac{x^j e^{-\frac{\eta x}{\chi_0}}}{j!} \right) * \left(\frac{x^{(j-1)} e^{-\frac{\varphi x}{\chi_0}}}{(j-1)!} \right) \right]. \tag{4.54}$$

By applying the Bessel function identity, if $\gamma = 2\frac{\sqrt{\tau c\theta\mu}}{\chi}$, and $\delta = \sqrt{\frac{\tau}{c\theta\mu}}$, we have

$$\exp\left\{ (\frac{\tau z}{\chi} + \frac{c\theta\mu}{\chi z})y \right\} = \sum_{k=-\infty}^{\infty} (\delta z)^k I_k(\gamma y).$$

Comparing the coefficients of z^k on both sides of Eq. (4.50) for $k = 0, 1, ..., N$, we obtain

$$\begin{aligned}
F_k(x) &= \frac{\eta}{\chi}\delta^k \int_0^x R(x-y) e^{\frac{-(\tau+\varphi+c\theta\mu)}{\chi} y} I_k(\gamma y) dy \\
&+ \frac{c\theta\mu}{\chi}\delta^k \int_0^x F_0(x-y) e^{\frac{-(\tau+\varphi+c\theta\mu)}{\chi} y} \left[I_k(\gamma y) - \delta I_{k+1}(\gamma y)\right] dy \\
&+ \frac{\tau}{\chi}\delta^k \int_0^x F_N(x-y) e^{\frac{-(\tau+\varphi+c\theta\mu)}{\chi} y} \\
&\left[\delta^{-N} I_{k-N}(\gamma u) - \delta^{-(N+1)} I_{k-1-N}(\gamma y)\right] dy \\
&+ \frac{\theta\mu}{\chi}\delta^k \int_0^x \sum_{m=1}^{c-1} (m-c) F_m(x-y) e^{\frac{-(\tau+\varphi+c\theta\mu)}{\chi} y} \\
&\left[\delta^{-(m-1)} I_{m-k-1}(\gamma y) - \delta^{-m} I_{m-k}(\gamma y)\right] dy. \qquad (4.55)
\end{aligned}$$

By setting $k = 0$ in Eq. (4.55), we have

$$\begin{aligned}
F_0(x) &= \frac{\eta}{\chi} \int_0^x R(x-y) e^{\frac{-(\tau+\varphi+c\theta\mu)}{\chi} y} I_0(\gamma y) dy \\
&+ \frac{c\theta\mu}{\chi} \int_0^x F_0(x-y) e^{\frac{-(\tau+\varphi+c\theta\mu)}{\chi} y} \left[I_0(\gamma y) - \delta I_1(\gamma y)\right] dy \\
&+ \frac{\tau}{\chi} \int_0^x F_N(x-y) e^{\frac{-(\tau+\varphi+c\theta\mu)}{\chi} y} \\
&\left[\delta^{-N} I_N(\gamma u) - \delta^{-(N+1)} I_{1+N}(\gamma y)\right] dy \\
&+ \frac{\theta\mu}{\chi} \int_0^x \sum_{m=1}^{c-1} (m-c) F_m(x-y) e^{\frac{-(\tau+\varphi+c\theta\mu)}{\chi} y} \\
&\left[\delta^{-(m-1)} I_{m-1}(\gamma y) - \delta^{-m} I_m(\gamma y)\right] dy. \qquad (4.56)
\end{aligned}$$

By applying Laplace transform to equation (4.56), we get

$$
\begin{aligned}
\left\{\frac{\varpi+\sqrt{\varpi^2-\gamma^2}}{2}-\frac{c\theta\mu}{\chi}\right\}F_0^*(s) &= \frac{\tau}{\chi}\left[\frac{\chi(\varpi-\sqrt{\varpi^2-\gamma^2})}{2\tau}\right]^N \\
&\left[1-\frac{\chi(\varpi-\sqrt{\varpi^2-\gamma^2})}{2\tau}\right]F_N^*(s) \\
&+\frac{\theta\mu}{\chi}\sum_{m=1}^{c-1}(m-c)F_m^*(s) \\
&\left[\frac{\chi(\varpi-\sqrt{\varpi^2-\gamma^2})}{2\tau}\right]^m \\
&\left[\frac{2\tau}{\chi(\varpi-\sqrt{\varpi^2-\gamma^2})}-1\right] \\
&+\frac{\eta}{\chi}R^*(s), \qquad (4.57)
\end{aligned}
$$

where $\varpi = s+\dfrac{\tau+c\theta\mu+\varphi}{\chi}$. By multiplying both sides of Eq. (4.57) by $\dfrac{\chi}{c\theta\mu}\left\{\dfrac{\chi(\varpi-\sqrt{\varpi^2-\gamma^2})}{2\tau}\right\}$, we simplify it to

$$
\begin{aligned}
\left\{1-\frac{\chi(\varpi-\sqrt{\varpi^2-\gamma^2})}{2\tau}\right\}F_0^*(s) &= \frac{\tau}{c\theta\mu}\left[\frac{r(\varpi-\sqrt{\varpi^2-\gamma^2})}{2\tau}\right]^{N+1} \\
&\left[1-\frac{\chi(\varpi-\sqrt{\varpi^2-\gamma^2})}{2\tau}\right]F_N^*(s) \\
&+\frac{1}{c}\sum_{m=1}^{c-1}(m-c)F_m^*(s) \\
&\left[\frac{\chi(\varpi-\sqrt{\varpi^2-\gamma^2})}{2\tau}\right]^m \\
&\left[1-\frac{\chi(\varpi-\sqrt{\varpi^2-\gamma^2})}{2\tau}\right] \\
&+\frac{\eta}{c\theta\mu}R^*(s) \\
&\left\{\frac{\chi(\varpi-\sqrt{\varpi^2-\gamma^2})}{2\tau}\right\}. \qquad (4.58)
\end{aligned}
$$

Solving Eq. (4.58) implies

$$\begin{aligned}
F_0^*(s) &= \frac{\tau}{c\theta\mu}\left[\frac{\chi(\varpi-\sqrt{\varpi^2-\gamma^2})}{2\tau}\right]^{N+1} F_N^*(s) \\
&+\frac{1}{c}\sum_{m=1}^{c-1}(m-c)F_m^*(s)\left[\frac{\chi(\varpi-\sqrt{\varpi^2-\gamma^2})}{2\tau}\right]^m \\
&+\frac{\eta}{c\theta\mu}R^*(s)\left[\frac{c\theta\mu}{\chi\left(s+\dfrac{\varphi}{\chi}\right)}\right. \\
&\left.-\frac{\tau}{\chi\left(s+\dfrac{\varphi}{\chi}\right)}\frac{\chi(\varpi-\sqrt{\varpi^2-\gamma^2})}{2\tau}\right]. \qquad (4.59)
\end{aligned}$$

Then, substituting $k = N$ in Eq. (4.55) yields to

$$\begin{aligned}
F_N(x) &= \frac{\eta}{\chi}\int_0^x R(x-y)e^{\dfrac{-(\tau+\varphi+c\theta\mu)}{\chi}y}\delta^N I_N(\gamma y)dy \\
&+\frac{c\theta\mu}{\chi}\int_0^x F_0(x-y)e^{\dfrac{-(\tau+\varphi+c\theta\mu)}{\chi}y} \\
&\left[\delta^N I_N(\gamma y)-\delta^{N+1}I_{N+1}(\gamma y)\right]dy \\
&+\frac{\tau}{\chi}\int_0^x F_N(x-y)e^{\dfrac{-(\tau+\varphi+c\theta\mu)}{\chi}y}\left[I_0(\gamma y)-\delta^{-1}I_1(\gamma y)\right]dy \\
&+\frac{\theta\mu}{\chi}\delta^N\int_0^x\sum_{m=1}^{c-1}(m-c)F_m(x-y)e^{\dfrac{-(\tau+\varphi+c\theta\mu)}{\chi}y} \\
&\left[\delta^{-(m-1)}I_{m-N-1}(\gamma y)-\delta^{-m}I_{m-N}(\gamma y)\right]dy. \qquad (4.60)
\end{aligned}$$

Applying Laplace transform of Eq. (4.60), we obtain

$$
\begin{aligned}
&\left[\sqrt{\varpi^2-\gamma^2}-\frac{\tau}{\chi}+\frac{\varpi-\sqrt{\varpi^2-\gamma^2}}{2}\right]F_N^*(s)-\frac{\tau}{\chi}\left[\frac{\chi(\varpi-\sqrt{\varpi^2-\gamma^2})}{2\tau}\right]^{N+1}\\
&\left[\frac{\chi(\varpi-\sqrt{\varpi^2-\gamma^2})}{2c\theta\mu}\right]^{N}\left[1-\frac{\chi(\varpi-\sqrt{\varpi^2-\gamma^2})}{2c\theta\mu}\right]F_N^*(s)=\\
&\frac{\eta}{\chi}R^*(s)\left[\frac{\chi(\varpi-\sqrt{\varpi^2-\gamma^2})}{2c\theta\mu}\right]^{N}\\
&\times\left\{1+\left[1-\frac{\chi(\varpi-\sqrt{\varpi^2-\gamma^2})}{2c\theta\mu}\right]\left[\frac{c\theta\mu}{\chi\left(s+\frac{\varphi}{\chi}\right)}-\frac{\tau}{\chi\left(s+\frac{\varphi}{\chi}\right)}\frac{\chi(\varpi-\sqrt{\varpi^2-\gamma^2})}{2\tau}\right]\right\}\\
&+\frac{\theta\mu}{\chi}\textstyle\sum_{m=1}^{c-1}(m-c)F_m^*(s)\left[\frac{\chi(\varpi-\sqrt{\varpi^2-\gamma^2})}{2\tau}\right]^{m}\\
&\times\left\{\left[1-\frac{\chi(\varpi-\sqrt{\varpi^2-\gamma^2})}{2c\theta\mu}\right]\left[\frac{\chi(\varpi-\sqrt{\varpi^2-\gamma^2})}{2c\theta\mu}\right]^{N}+\left[\frac{2\tau}{\chi(\varpi-\sqrt{\varpi^2-\gamma^2})}\right]^{N}\right.\\
&\left.\times\left[\frac{2\tau}{\chi(\varpi-\sqrt{\varpi^2-\gamma^2})}-1\right]\right\}.
\end{aligned}
\tag{4.61}
$$

Further, Eq. (4.61) can be rewritten as:

$$
(1-f^*(s))F_N^*(s)=G^*(s)+K^*(s), \tag{4.62}
$$

where

$$
\begin{aligned}
f^*(s) &= \left[\frac{\chi(\varpi-\sqrt{\varpi^2-\gamma^2})}{2c\theta\mu}\right]^{N+1}\left[\frac{\chi(\varpi-\sqrt{\varpi^2-\gamma^2})}{2\tau}\right]^{N+1}\\
&\quad\left[1-\frac{\chi(\varpi-\sqrt{\varpi^2-\gamma^2})}{2c\theta\mu}\right]\\
&\quad+\left[\frac{\varpi-\sqrt{\varpi^2-\gamma^2}}{2c\theta\mu}\right],
\end{aligned}
\tag{4.63}
$$

$$
\begin{aligned}
G^*(s) &= \frac{\eta}{\tau}R^*(s)\left[\frac{\chi(\varpi-\sqrt{\varpi^2-\gamma^2})}{2c\theta\mu}\right]^{N+1}\\
&\quad\left[1+\left[1-\frac{\chi(\varpi-\sqrt{\varpi^2-\gamma^2})}{2c\theta\mu}\right]\right.\\
&\quad\left.\left[\frac{c\theta\mu}{\chi\left(s+\frac{\varphi}{\chi}\right)}-\frac{\tau}{\chi\left(s+\frac{\varphi}{\chi}\right)}\frac{\chi(\varpi-\sqrt{\varpi^2-\gamma^2})}{2\tau}\right]\right],
\end{aligned}
\tag{4.64}
$$

and

$$K^*(s) = \frac{\theta\mu}{\tau}\sum_{m=1}^{c-1}(m-c)F_m^*(s)\left[\frac{\chi(\varpi-\sqrt{\varpi^2-\gamma^2})}{2\tau}\right]^m \\ \left[\left[1-\frac{\chi(\varpi-\sqrt{\varpi^2-\gamma^2})}{2c\theta\mu}\right]\left[\frac{\chi(\varpi-\sqrt{\varpi^2-\gamma^2})}{2c\theta\mu}\right]^{N+1} \right.\\ +\left[\frac{\chi(\varpi-\sqrt{\varpi^2-\gamma^2})}{2c\theta\mu}\right]\left[\frac{2\tau}{\chi(\varpi-\sqrt{\varpi^2-\gamma^2})}\right]^N \\ \left.\left[\frac{2\tau}{\chi(\varpi-\sqrt{\varpi^2-\gamma^2})}-1\right]\right]. \quad (4.65)$$

Equation (4.62) can then be expressed as:

$$F_N^*(s) = [G^*(s)+K^*(s)]\sum_{i=0}^{\infty}[f^*(s)]^i. \quad (4.66)$$

On inversion, we get

$$F_N(x) = [G(x)+K(x)] * \sum_{i=0}^{\infty}[f(x)]^{*i}, \quad (4.67)$$

where

$$f(x) = e^{-\left(\frac{\tau+\varphi+c\theta\mu}{\chi}\right)x}\left[\frac{1}{2c\theta\mu}I_1(\gamma x)\gamma + (2N+2)\frac{I_{2N+2}(\gamma x)}{x} - \sqrt{\frac{\tau}{c\theta\mu}}(2N+3)\frac{I_{2N+3}(\gamma x)}{x}\right],$$

$$K(x) = \frac{\theta\mu}{\tau}\sum_{m=1}^{c-1}(m-c)F_m(x)\frac{1}{x}e^{-\left(\frac{\tau+\varphi+c\theta\mu}{\chi}\right)x} \\ *\left[\delta^{N+1-m}(N+m+1)I_{N+m+1}(\gamma x)\right. \\ -\delta^{N+2-m}(N+m+2)I_{N+m+2}(\gamma x) \\ +\delta^{N+2-m}(m-N)I_{m-N}(\gamma x) \\ \left.-\delta^{N+1-m}(m+1-N)I_{m+1-N}(\gamma x)\right], \quad (4.68)$$

and

$$G(x) = \frac{\eta}{\tau}R(x) * h(x),$$

such that

$$h(x) = \left(\frac{\tau}{c\theta\mu}\right)^{\frac{N+1}{2}} \frac{(N+1)I_{N+1}(\gamma x)}{x} e^{-\left(\frac{\tau+\varphi+c\theta\mu}{\chi}\right)x}$$
$$+e^{-\frac{\varphi x}{\chi}} * e^{-(\frac{\tau+\varphi+c\theta\mu}{\chi})x} \frac{\sqrt{\tau}^{N+1}}{\chi\sqrt{c\theta\mu}^{N+1}}(N+1)\frac{I_{N+1}(\gamma x)}{x}$$
$$-\frac{2\tau}{\chi}\sqrt{\frac{\tau}{c\theta\mu}}^{N}(N+2)\frac{I_{N+2}(\gamma x)}{x}$$
$$+\frac{\tau}{\chi}\sqrt{\frac{\tau}{c\theta\mu}}^{N+1}(N+3)\frac{I_{N+3}(\gamma x)}{x}. \quad (4.69)$$

Then, substituting Eq. (4.66) into (4.59) and after simplification, we obtain

$$F_0^*(s) = \frac{\eta}{c\theta\mu}\Theta_1(s)R^*(s) + \Theta_2(s), \quad (4.70)$$

where

$$\Theta_1(s) = \frac{c\theta\mu}{\chi\left(s+\frac{\varphi}{\chi}\right)} - \frac{\tau}{\chi\left(s+\frac{\varphi}{\chi}\right)}\frac{\chi(\varpi-\sqrt{\varpi^2-\gamma^2})}{2\tau} + \sum_{i=0}^{\infty}[f^*(s)]^i$$
$$\left[\frac{\left[\chi(\varpi-\sqrt{\varpi^2-\gamma^2})\right]^{2N+2}}{(2\tau)^{N+1}(2c\theta\mu)^{N+1}} + \frac{c\theta\mu}{(\chi s+\varphi)}\frac{\left[\chi(\varpi-\sqrt{\varpi^2-\gamma^2})\right]^{2N+2}}{(2\tau)^{N+1}(2c\theta\mu)^{N+1}}\right.$$
$$\left.-\frac{2}{(\chi s+\varphi)}\frac{\left[\chi(\varpi-\sqrt{\varpi^2-\gamma^2})\right]^{2N+3}}{(2\tau)^{N+1}(2c\theta\mu)^{N+1}} + \frac{\tau}{(\chi s+\varphi)}\frac{\left[\chi(\varpi-\sqrt{\varpi^2-\gamma^2})\right]^{2N+4}}{(2\tau)^{N+2}(2c\theta\mu)^{N+2}}\right],$$

$$\Theta_2(s) = \frac{1}{c}\sum_{m=1}^{c-1}(m-c)F_m^*(s)\left[1+\sum_{i=0}^{\infty}[f^*(s)]^i\left[\frac{\left[\chi(\varpi-\sqrt{\varpi^2-\gamma^2})\right]^{2N+2+m}}{(2\tau)^{N+1+m}(2c\theta\mu)^{N+1}}\right.\right.$$
$$-\frac{\left[\chi(\varpi-\sqrt{\varpi^2-\gamma^2})\right]^{2N+3+m}}{(2\tau)^{N+1+m}(2c\theta\mu)^{N+2}} + \frac{\left[\chi(\varpi-\sqrt{\varpi^2-\gamma^2})\right]^{m+1}}{2(2\tau)^m c\theta\mu}$$
$$\left.\left.-\frac{\left[\chi(\varpi-\sqrt{\varpi^2-\gamma^2})\right]^{m+2}}{2(2\tau)^{m+1}c\theta\mu}\right]\right]$$

which on inversion yields

$$F(x) = \frac{\eta}{c\theta\mu}\Theta_1 R(x) + \Theta_2, \quad (4.71)$$

where

$$\Theta_1 = \frac{c\theta\mu}{\chi} e^{-\frac{\varphi}{\chi}x} - \frac{\sqrt{\tau c\theta\mu} I_1(\gamma x)}{\chi x} e^{-\left(\frac{\tau + c\theta\mu + \varphi}{\chi}\right)x} * e^{-\frac{\varphi}{\chi}x}$$

$$+ \sum_{i=0}^{\infty} [f(x)]^{*i} * \left[(2N+2) I_{2N+2}(\gamma x) \frac{e^{-\left(\frac{\tau + c\theta\mu + \varphi}{r}\right)x}}{x} \right.$$

$$+ \frac{c\theta\mu}{\chi}(2N+2) I_{2N+2}(\gamma x) \frac{e^{-\left(\frac{\tau + c\theta\mu + \varphi}{\chi}\right)x}}{x} * e^{-\frac{\varphi}{\chi}x}$$

$$- \frac{2\sqrt{\tau c\theta\mu}}{\chi}(2N+3) I_{2N+3}(\gamma x) \frac{e^{-\left(\frac{\tau + c\theta\mu + \varphi}{\chi}\right)x}}{x} * e^{-\frac{\varphi}{\chi}x}$$

$$\left. + \frac{\tau}{\chi}(2N+4) I_{2N+4}(\gamma x) \frac{e^{-\left(\frac{\tau + c\theta\mu + \varphi}{\chi}\right)x}}{x} * e^{-\frac{\varphi}{\chi}x} \right],$$

and

$$\begin{aligned} \Theta_2 &= \frac{1}{c} \sum_{m=1}^{c-1} (m-c) F_m(x) * e^{-\left(\frac{\tau + c\theta\mu + \varphi}{\chi}\right)x} \\ & \left[1 + \sum_{i=0}^{\infty} [f(x)]^{*i} \left[\sqrt{\frac{c\theta\mu}{\tau}}^{m} \frac{(2N+2+m) I_{2N+2+m}(\gamma x)}{x} \right.\right. \\ & - \sqrt{\frac{c\theta\mu}{\tau}}^{m-1} \frac{(2N+3+m) I_{2N+3+m}(\gamma x)}{x} + \sqrt{\frac{c\theta\mu}{\tau}}^{m-1} \frac{(m+1) I_{m+1}(\gamma x)}{x} \\ & \left.\left. - \sqrt{\frac{c\theta\mu}{\tau}}^{m} \frac{(m+2) I_{m+2}(\gamma x)}{x} \right]\right]. \end{aligned}$$

From equation (4.54), we can see that $R(x)$ has an explicit expression. Moreover, Eqs. (4.71)-(4.67) give the expressions for $F_0(x)$ and $F_N(x)$ in terms of $R(x)$.
Similarly, for $k = 1, 2, ..., N-1$, Eq. (4.55) shows that $F_k(x)$ depends on $F_0(x)$, $F_N(x)$ and $R(x)$, which are all known. Therefore, we can explicitly find all the joint stationary probabilities for the fluid model.

4.5.2 Analyzing the stationary buffer content distribution

The fluid model under consideration has a stationary buffer content distribution given by

$$F(x) = P\{C \leqslant x\} = R(x) + \sum_{k=0}^{N} F_k(x).$$

Applying Laplace transform to both sides of the equation, we get

$$\begin{aligned} F^*(s) &= R^*(s) + \sum_{k=0}^{N} F_k^*(s) \\ &= R^*(s) + \frac{\eta}{\chi s + \varphi} R^*(s), \end{aligned} \tag{4.72}$$

which can be inverted to give

$$F(x) = R(x) + \frac{\eta}{\chi} e^{-\frac{\varphi}{\chi} x} * R(x), \tag{4.73}$$

where $R(x)$ is defined by equation (4.54). We still need to find the constant $'a'$. The constant $'a'$ can be obtained using the fact

$$\lim_{s \longrightarrow 0} sF^*(s) = 1. \tag{4.74}$$

Using equation (4.53 to substitute for $R^*(s)$ in $F^*(s)$ and then using equation (4.74), we have

$$a = \frac{\chi_0 \varphi + \chi \eta}{\chi_0(\eta + \varphi)}.$$

4.5.3 Validation of analytical expressions for fluid model

We compare the analytical expressions for $F_0(x)$ and $F_k(x)$ in the Laplace domain with the existing results. When $N \longrightarrow \infty$, the model becomes a fluid queue driven by an $M/M/1$ queue with disaster and repair. We can see from Eq. (4.53) that

$$R^*(s) = \frac{a}{s} - \frac{a\eta/\chi_0}{s(s + \frac{\varphi}{\chi} + \frac{\eta}{\chi_0})}.$$

Also, as $N \longrightarrow \infty$, from equation (4.59), we have

$$\begin{aligned}
F_0^*(s) &= \frac{\eta}{\chi} R^*(s) \left[\frac{c\theta\mu}{\chi s + \varphi} - \frac{\chi(\varpi - \sqrt{\varpi^2 - \gamma^2})}{2\,(s\chi + \varphi)} \right] \\
&\quad + \frac{1}{c} \sum_{m=1}^{c-1} (m-c) F_m^*(s) \left[\frac{\chi(\varpi - \sqrt{\varpi^2 - \gamma^2})}{2\tau} \right]^m \\
&= \frac{\eta}{\chi} \left[\frac{c\theta\mu}{\chi s + \varphi} - \frac{f(s)}{\chi s + \varphi} \right] R^*(s) \\
&\quad + \frac{1}{c} \sum_{m=1}^{c-1} (m-c) F_m^*(s) \left[\frac{\chi(\varpi - \sqrt{\varpi^2 - \gamma^2})}{2\tau} \right]^m \\
&= \frac{a\eta(c\theta\mu - f(s))}{c\theta\mu\chi s \left(s + \dfrac{\varphi}{\chi} + \dfrac{\eta}{\chi_0} \right)} \\
&\quad + \frac{1}{c} \sum_{m=1}^{c-1} (m-c) F_m^*(s) \left[\frac{r(\varpi - \sqrt{\varpi^2 - \gamma^2})}{2\tau} \right]^m, \qquad (4.75)
\end{aligned}$$

where $f(s) = \dfrac{\chi(\varpi - \sqrt{\varpi^2 - \gamma^2})}{2}$. In the same way, Laplace transform of equation (4.55) as $N \longrightarrow \infty$, gives

$$\begin{aligned}
F_k^*(s) &= \frac{\eta\gamma^{-k}\delta^k[\varpi - \sqrt{\varpi^2 - \gamma^2}]^k}{\chi\sqrt{\varpi^2 - \gamma^2}} R^*(s) + \frac{c\theta\mu\gamma^{-k}\delta^k[\varpi - \sqrt{\varpi^2 - \gamma^2}]^k}{\chi\sqrt{\varpi^2 - \gamma^2}} F_0^*(s) \\
&\quad - \frac{c\theta\mu\gamma^{-(k+1)}\delta^{k+1}[\varpi - \sqrt{\varpi^2 - \gamma^2}]^k}{\chi\sqrt{\varpi^2 - \gamma^2}} F_0^*(s) \\
&\quad + \frac{\theta\mu}{\chi\sqrt{\varpi^2 - \gamma^2}} \sum_{m=1}^{c-1} (m-c) F_m^*(s) \left[\frac{\chi(\varpi - \sqrt{\varpi^2 - \gamma^2})}{2\tau} \right]^{m-k} \left[\frac{2\tau}{\chi(\varpi - \sqrt{\varpi^2 - \gamma^2})} - 1 \right] \\
&= \frac{\chi^{k-1}[\varpi - \sqrt{\varpi^2 - \gamma^2}]^k}{2^k(c\theta\mu)^{k-1}\sqrt{\varpi^2 - \gamma^2}} \left[\frac{2\tau}{\chi(\varpi - \sqrt{\varpi^2 - \gamma^2})} - \frac{\chi(\varpi - \sqrt{\varpi^2 - \gamma^2})}{2c\theta\mu} \right] F_0^*(s) \\
&\quad + \frac{\theta\mu}{\chi\sqrt{\varpi^2 - \gamma^2}} \sum_{m=1}^{c-1} (m-c) F_m^*(s) \left[\frac{\chi(\varpi - \sqrt{\varpi^2 - \gamma^2})}{2\tau} \right]^{m-k} \left[\frac{2\tau}{r(\varpi - \sqrt{\varpi^2 - \gamma^2})} - 1 \right] \\
&= \frac{\chi^k}{(2c\theta\mu)^k} [\varpi - \sqrt{\varpi^2 - \gamma^2}]^k F_0^*(s) \\
&\quad + \frac{\theta\mu}{\chi\sqrt{\varpi^2 - \gamma^2}} \sum_{m=1}^{c-1} (m-c) F_m^*(s) \left[\frac{\chi(\varpi - \sqrt{\varpi^2 - \gamma^2})}{2\tau} \right]^{m-k} \left[\frac{2\tau}{\chi(\varpi - \sqrt{\varpi^2 - \gamma^2})} - 1 \right].
\end{aligned}$$

4.6 Conclusion

In this chapter, we studied a $M/M/c/N$ queueing model with Bernoulli feedback, catastrophes and repairs. We derived the transient and steady-state solutions for this model and obtained the expression for the time-dependent and time-independent expected number of customers in the system. We have also modeled the system as a fluid queue and analyzed its stationary solution.

Our research is motivated by several factors, such as the complexity and importance of multi-server queuing models, the dynamics of recurring customers in the system, the relevance of disasters and repairs in different applications, the design of system performance based on transient and steady state behavior, and the understanding of system behavior through the fluid queue model.

We showed that our model can capture the effects of disasters and repairs on the system performance and reliability, and provide insights into the optimal solutions for improving the system capacity and resilience. We also demonstrated the usefulness and applicability of our model for various real-world systems, such as call centers, hospitals, and manufacturing facilities. Our paper contributes to the field of queueing theory by providing a novel and comprehensive analysis of a multi-server queueing model with different features.

Some possible directions for further extensions of queueing models with catastrophes and repairs are:

1. Considering different types of catastrophes, such as partial or complete, that affect only a fraction or all of the customers in the system.

2. Studying the impact of retrial, or impatience on the system performance and reliability.

3. Analyzing the optimal control policies for minimizing the expected cost or maximizing the expected profit of the system.

Conclusion

This thesis has presented a comprehensive exploration of complex queueing systems modeling various aspects of machining operations under realistic conditions. Through three distinct research works, we have delved into the intricate dynamics of these systems, employing advanced analytical techniques to derive important performance measures and optimize system parameters.

Our first study on the infinite-capacity multi-server queueing system with batch arrivals, Bernoulli feedback, working breakdowns, balking, and reneging has provided valuable insights into system behavior under extreme conditions. The use of probability generating functions and supplementary variable techniques allowed us to derive stability conditions and steady-state probabilities, leading to the development of a cost-effective model for economic analysis.

The second research work on the finite-capacity unreliable multi-server queueing system with working vacations, Bernoulli interruptions, and threshold-based recovery policy has deepened our understanding of system performance under varied operational conditions. The application of matrix-analytic methods, particularly the quasi-birth-death process approach, enabled us to derive steady-state solutions and explore cost parameter optimization.

Our third study on the finite-capacity M/M/c/N queueing system with Bernoulli feedback, catastrophes, and repairs has shed light on the impact of system failures on time-dependent performance measures and busy periods. The combination of Markovian analysis and probabilistic approaches provided insights applicable to finite source systems such as call centers, manufacturing facilities, and service organizations.

Implications

The findings from these studies have several important implications for both theory and practice:

1. System Design: Our models provide a framework for designing more robust and efficient machining systems that can better handle uncertainties and disruptions.

2. Performance Optimization: The derived performance measures and

optimization techniques can guide managers in fine-tuning system parameters to achieve optimal performance under various conditions.

3. Economic Analysis: The cost models developed in our studies can assist in making informed decisions about system investments and operational strategies.

4. Reliability Engineering: Our findings contribute to the field of reliability engineering by providing insights into system behavior under failures and repair scenarios.

5. Customer Service: The incorporation of customer behavior (balking, reneging, feedback) in our models can inform strategies for improving customer satisfaction in service-oriented systems.

Further extensions

While this thesis has made significant contributions to the field, there are several avenues for future research:

1. Non-Markovian models: Extending the current models to incorporate non-exponential distributions for arrival, service, and repair times could provide more realistic representations of certain systems.

2. Multi-class customers: Investigating systems with multiple classes of customers with different priorities and service requirements could offer insights into more complex real-world scenarios.

3. Network models: Expanding the single-node models to queueing networks could help in analyzing more complex manufacturing or service systems.

4. Machine Learning integration: Incorporating machine learning techniques for dynamic parameter adjustment and predictive maintenance could improve the adaptability and efficiency of these systems.

5. Fluid and diffusion approximations: Developing fluid and diffusion approximations for these models could provide simpler analytical tools for performance analysis, especially for large-scale systems.

6. Optimal control: Investigating optimal control policies for these complex systems, particularly in the presence of time-varying parameters or state-dependent control.

Bibliography

[1] Ahmadi M., Golkarifard M., Movaghar A., Yousefi H., Processor sharing queues with impatient customers and state-dependent rates, IEEE/ACM Transactions on Networking, 29(6), (2021), 2467-2477.

[2] Ahuja A., Jain A., 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),(2023), 7499-7518.

[3] Altman E., Yechiali U., Infinite-server queues with system's additional tasks and impatient customers, Probability in the Engineering and Informational Sciences, 22(4), (2008), 477-493.

[4] Ammar S. I., Transient analysis of an $M/M/1$ queue with impatient behavior and multiple vacations, Appl. Math. Comp., 260, (2015), 97-105.

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

[6] Ammar S. I., Jiang T., Qingqing Y., Transient analysis of impatient customers in an $M/M/1$ disasters queue in random environment, Engineering Computations, 37(6), (2020), 1945-1965.

[7] Avi-Itzhak B., Naor P., Some queuing problems with the service station subject to breakdown, Operations Research 11(3),(1963), 303-320.

[8] Ayyappan G., Karpagam S., An $M^{[X]}/G(a,b)/1$ queueing system with breakdown and repair, standby server, multiple vacation and control policy on request for re-service, Mathematics, 6(6),(2018), 1-18. http://www.mdpi.com/2227–7390/6/6/101.

[9] Ayyappan G., Muthu Ganapathi Subramanian A., Sekar G., $M/M/1$ retrial queuing system with loss and feedback under non-pre-emptive priority service by matrix geometric method, Applied Math. Sci., 4(48), (2010), 2379-2389.

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

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

[12] Baccelli F., Brémaud P., Elements of queueing theory: Palm Martingale calculus and stochastic recurrences, Springer Science Business Media, (26),(2002).

[13] Bank B., Samanta S. K., Analytical and computational studies of the $BMAP/G(a,Y)/1$ queue. Communications in Statistics-Theory and Methods, 50(15), (2021), 3586-3614.

[14] Barbhuiya F. P., Kumar N., Gupta U. C., Analysis of the $GI/M/c$ queue with N-threshold policy, Quality Technology Quantitative Management, 19(4), (2022), 490-510.

[15] Bekker R., Mandjes M., A fluid model for a relay node in an ad hoc network: the case of heavy-tailed input, Mathematical Methods of Operations Research, 70(2),(2009), 357-384.

[16] Bhat U. N., An Introduction to queueing theory, modling and analysis in application, Statistics for Industry and Technology, Birkhauser, Boston, (2008).

[17] Bonomi F., Mitra D., Seery J., Adaptive algorithms for feedback-based below control in highspeed wide-area ATM network, IEEE J. Sel. Areas Commun, 13(7),(1995), 1267-1283.

[18] Bosman J. W., Nunez-Queija R., A spectral theory approach for extreme value analysis in a tandem of fluid queues, Queueing Systems, Theory and Applications, 78(2), (2014), 121-154.

[19] Boualem M., Stochastic analysis of a single server unreliable queue with balking and general retrial time, Discrete and continuous models and applied computational science, 28(4),(2020), 319–326.

[20] Bouchentouf A. A., Belarbi F., Performance evaluation of two Markovian retrial queuing model with balking and feedback, Acta Univ. Sapientiae, Mathematica, 5, (2013), 132-146.

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

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

[23] Bouchentouf A. A., Cherfaoui M., Boualem M., Performance and economic analysis of a single server feedback queueing model with vacation and impatient customers, opsearch , 56(1), (2019), 300-320.

[24] Bouchentouf A. A., Guendouzi A., Cost optimization analysis for an $M^X/M/c$ vacation queueing system with servers and impatient customers, SeMA Journal,76(2), (2019), 309–341.

[25] Bouchentouf A. A., Guendouzi A., Single server batch arrival Bernoulli feedback queueing system with waiting server, K-Variant Vacations and Impatient Customers, Operations Research Forum, 2(1), (2021), 1–23.

[26] Bouchentouf A. A., Guendouzi A. H., The $M^X/M/c$ bernoulli feedback queue with variant multiple working vacations and impatient customers performance and economic analysis, Arabian Journal Of Mathematics, 9(2), (2020), 309-327.

[27] Bouchentouf A. A., Kadi K., Rabhi A., Analysis of two heterogeneous server queueing model with balking, reneging and feedback, Math.Sci. Appl. E Notes, 2(2), (2014), 10-21.

[28] Bouchentouf A. A., Yahiaoui L., On feedback queueing system with reneging and retention of reneged customers, multiple working vacations and Bernoulli schedule vacation interruption. Arabian Journal of Mathematics, 6(1), (2017), 1-11.

[29] Bouchentouf A. A., Yahiaoui L., Kadi M., Majid S., Impatient customers in Markovian queue with Bernoulli feedback and waiting server under variant working vacation policy, Operations Research and Decisions, 30(4), (2020), 5-28.

[30] Boudali O., Economou A., Optimal and equilibrium balking strategies in the single server Markovian queue with catastrophes, European Journal of Operational Research, 218(3), (2012), 708–715.

[31] Boudali O., Economou A., The effect of catastrophes on the strategic customer behavior in queueing systems, Naval Research Logistics, 60(7), (2013), 571–587.

[32] Brémaud P., Markov chains, Springer Verlag, (1999).

[33] Brown M., Ross S. M., Some results for infinite server Poisson queues, Journal of Applied Probability, 6(3), (1969), 604–611.

[34] Cao J., Cheng K., Analysis of $M/G/1$ queueing system with repairable service station, Acta Math. Appl. Sinica, 5, (1982), 113–127.

[35] Chakravarthy S. R., Kulshrestha R., A queueing model with server breakdowns, repairs, vacations, and backup server, Operations research perspectives, 7, (2020), 100131.

[36] Chakravarthy S. R., Shruti and Rumyantsev A., Analysis of a queueing model with batch Markovian arrival process and general distribution for group clearance, Methodol Comput Appl Probab., 23, (2021), 1551–1579.

[37] Chao X., A queuing network model with catastrophes and product form solution, Operations Research Letters, 18(2), (1995), 75-79.

[38] Chang F. M., Lee Y. T., Chang C. J., and Yeh C., Analysis of a standby redundant system with controlled arrival of failed machines, International Journal of Industrial and Systems Engineering, 28(1), (2018), 117–134.

[39] Chen W. L., and Wang K. H., Reliability analysis of a retrial machine repair problem with warm standbys and a single server with N-policy, Reliab. Eng. Syst. Saf, 180,(2018), 476–486.

[40] Cherfaoui M., Bouchentouf A. A., Boualem M., Modelling and simulation of Bernoulli feedback queue with general customers' impatience under variant vacation policy, *IJOR*, 46(4),(2023), 451-480.

[41] Chiamsiri S., Leonard M.S., A diffusion approximation for bulk queues, Management Science, 27 (10), (1981), 1188–1199.

[42] Cinlar E., Introduction to Stochastic Processes, Prentice Hall, (1975).

[43] Cox D. R, Some statistical methods connected with series of events, Journal of the Royal Statistical Society, Series B (Methodological), 17(2), (1955), 129-157.

[44] Crescenzo Di A., Giorno V., Nobile A. G., On the $M/M/1$ queue with catastrophes and its continuous approximation, Queueing Systems, 43, (2003), 329-347.

[45] Dimou S., Economou A., The single server queue with catastrophes and geometric reneging, Methodology & Computing in Applied Probability, 15(3), (2013), 595–621.

[46] Dombacher, C., Queueing models for call centres, Nikolaus Lenaugasse 8, A-2232 Deutsch-Wagram, (2010).

[47] Dong J., Rouba I., SRPT scheduling discipline in many-server queues with impatient customers, Management Science, 67(12), (2021), 7708-7718.

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

[49] Doshi B. T., Single sever queues with vacations, In Stochastic analysis of computer and communication systems, ed. H. Takagi, Amsterdam: North Holland, (1990), 217–65.

[50] Dudin S. A., Dudina O. S., Kostyukova O. I., Analysis of a Queuing System with possibility of waiting customers jockeying between two groups of servers, Mathematics, 11(1475), (2023), 1-21.

[51] Foster FG., Batched queuing processes, Operations Research, 12(3), (1964), 441-449.

[52] Gao S., Z. Liu Z., An $M/G/1$ queue with single working vacation and vacation interruption under Bernoulli schedule, Appl. Math. Model, 37(3), (2013), 1564–1579.

[53] Goswami V., Relationship between randomized F-policy and randomized N-policy in discrete-time queues, OP, 53(1), (2015), 131-150.

[54] Goswami V., Mund G. B., Analysis of renewal input batch service queue with impatient customers and multiple working vacations, International Journal of Management Science and Engineering Management, 15(2), (2020), 96-105.

[55] Goswami C., Selvaraju N., A working vacation queue with priority customers and vacation interruptions, International Journal of Operational Research, 17(3), (2013), 311-332.

[56] Graaf W.F., Scheinhardt W.R.W. and Boucherie R.J., Shot-noise fluid queues and infinite-server systems with batch arrivals, Performance evaluation, 116, (2017), 143–155.

[57] Gray W. J., Wang P. P., Scott M. K., A queueing model with multiple types of server breakdowns, Quality Technology Quantitative Management, 1(2), (2004), 245-255.

[58] Grill K., Theorie stochastischer prozesse, TU Wien, (2007).

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

[60] Hanukov G., Hassoun M., Oren M., On the benefits of providing timely information in ticket queues with balking and calling times, Mathematics, 9(2753), (2021), 1-16.

[61] He X., Hu W., Modeling relief demands in an emergency supply chain system under large-scale disasters based on a queuing network, The Scientific World Journal,Volume 2014, Article ID 195053, 12 pages http://dx.doi.org/10.1155/2014/195053

[62] Huang Y. H., Zhang T. C., Reduction of nitrobenzene and formation of corrosion coatings in zerovalent iron systems, Water Research, 40(16), (2006), 3075-3082.

[63] Jain M., Priority queue with batch arrival, balking, threshold recovery, unreliable server and optimal service, Rairo-Oper. Res, 51(2), (2017), 417–432.

[64] Jain M., Jain A., Working vacations queueing model with multiple types of server breakdowns, Appl. Math. Model, 1(13), (2010), 34.

[65] Jain J. L., Mohanty S. G., Böhm W., A course on queueing models, CRC Press,(2006).

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

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

[68] Jayachitra P., Albert A. J. , Recent developments in queueing models under N-policy, A short survey. Int. J. Math. Arch, 5 (3), (2014), 227–233.

[69] Jiang T., Liu L., The $GI/M/1$ queue in a multi-phase service environment with disasters and working breakdowns, Int. J. Comput. Math, 94, (2017), 707–726.

[70] Jiang T., Liu L., Li J., Analysis of the $M/G/1$ queue in multi-phase random environment with disasters, Journal of Mathematical Analysis & Applications, 430(2), (2015), 857–873.

[71] Jolai F., Asadzadeh S. M., Taghizadeh M. R., Performance estimation of an email contact center by a finite source discrete time $Geo/Geo/1$ queue with disasters, Computers & Industrial Engineering, 55, (2008), 543–556.

[72] Kalidass K., Ramanath K., Time dependent analysis of $M/M/1$ queue with server vacations and a waiting server, In Proceedings of the 6th International Conference on Queueing Theory and Network Applications, (2011), 77-83.

[73] Kim B. K., Bueker O., Non-preemptive priority $M/M/m$ queue with servers' vacations, Computers Industrial Engineering, 160,(2021), 107390.

[74] Kim B. K., Lee D. H., The $M/G/1$ queue with disasters and working breakdowns, Appl. Math. Model, 38, (2014), 1788–1798.

[75] Koroliuk V. S., Melikov A. Z., Ponomarenko L. A., Rustamov A. M., Methods for analysis of multi-channel queueing system with instantaneous and delayed feedbacks, Cybern. Syst. Analysis, 52, (2016), 58-70.

[76] Kulkarni V. G., Yan K., A fluid model with upward jumps at the boundary, Queueing Systems, 56(2), (2007), 103-117.

[77] Kumar K. B., Arivudainambi D., Transient solution of an $M/M/1$ queue with catastrophes, Computers & Mathematics with Applications, 40(10), (2000), 1233–1240.

[78] Kumar K. B., Arivudainambi D., Vijayakumar A., An $M/G/1/1$ queue with unreliable server and no waiting capacity, International journal of information and management sciences, 13(2), (2002), 35-50.

[79] Kumar R., Sapana S., Transient analysis of an $M/M/c$ queuing system with retention of reneging customers, International Journal of Operational Research, 36(1), (2019), 78-91.

[80] Kumar R., Soodan B. S., Kuaban G. S., Czekalski P., Sharma, S., Performance analysis of a cloud computing system using queuing model with correlated task reneging, In Journal of Physics: Conference Series, 2091(1), (2021).

[81] Kumari U., Sharma D. C., Performance analysis of a warm standby machine repair problem with servers vacation, impatient and controlling F-policy, Mathematics in Engineering, science and aerospace, 12(2), (2021), 1–21.

[82] Latouche G., Taylor P. G., A stochastic fluid model for an ad hoc mobile network, Queueing Systems, 63(1-4), (2009), 109-129.

[83] Lee D. H. , Optimal pricing strategies and customers' equilibrium behavior in an unobservable $M/M/1$ queueing system with negative customers and repair, Mathematical Problems in Engineering, Volume 2017, Article ID 8910819, 11 pages https://doi.org/10.1155/2017/8910819

[84] Lee S., Dudin A., Dudina O., Kim C., Analysis of a priority queueing system with the enhanced fairness of servers scheduling, Journal of Ambient Intelligence and Humanized Computing, (2022), 1-13.

[85] Lee S., Lee H.W., Yoon S.H., Chae K.C., Batch arrival queue with N-policy and single vacation, Computers & Operations Research, 22(2), (1995), 173–189.

[86] Li Q., Guo P., Wang Y. , Equilibrium analysis of unobservable $M/M/n$ priority queues with balking and homogeneous customers, Operations Research Letters, 48(5), (2020), 674-681.

[87] Liu L., Templeton J.GC., Autocorrelations in infinite server batch arrival queues, Queueing Syst, 14(3–4), (1993), 313–337.

[88] Lucantoni D. M., New results on the single server queue with a batch Markovian arrival process, Communications in Statistics. Stochastic Models, 7(1), (1991), 1–46.

[89] Lv S., Li J., The $M/M/N$ repairable queueing system with variable breakdown rates, Discrete Dynamics in Nature and Society, 2013,(2013), 10.

[90] Ma Z., Cui G., Wang P., Hao Y.,$M/M/1$ vacation queueing system with working breakdowns and variable arrival rate, J. Comput. Inf. Syst, 11, (2015), 1545–1552.

[91] Ma S., Guo W., Strategic joining rules in unobservable queues with dynamic service rate, IMA Journal of Management Mathematics, 34(4), (2023), 779-801.

[92] Mahanta S., Kumar N., Choudhury G., An analytical approach of Markov modulated Poisson input with feedback queue and repeated service under N-policy with setup time, Quality Technology Quantitative Management, (2023), 1-29.

[93] Masuyama H., Takine T., Analysis of an infinite-server queue with batch Markovian arrival streams, Queueing Syst., 42(3), (2002), 269–296.

[94] Melikov A., Aliyeva S., Refined approximate algorithm for steady-state probabilities of the large scale queuing systems with instantaneous and delayed feedback, in: A. Dudin, A. Nazarov, A. Moiseev (eds.), Inform. Technol. and Math. Modelling, Queueing Theory and Applications, ITMM 2019, Communications in Computer and Information Science, 1109, (2019), 188-201.

[95] Melikov A., Aliyeva S., Nair S. S., Kumar B. K., Retrial queuing-inventory systems with delayed feedback and instantaneous damaging of items, Axioms, 11(5), (2022),241.

[96] Melikov A. Z., Zadiranova L., Moiseev A., Two asymptotic conditions in queue with MMPP arrivals and feedback, Communications in Computer and Inform. Sci., 678, (2016), 231-240.

[97] Mitra D., Stochastic theory of a fluid model of producers and consumers coupled by a buffer, Advances in Applied Probability, 20(3), (1988), 646-676.

[98] Mytalas G. C., Zazanis M. A., An $M^X/G/1$ queueing system with disasters and repairs under a multiple adapted vacation policy, Naval Research Logistics, 62(3), (2015), 171–189.

[99] Nagarajan R., Kurose J., On defining, computing and guaranteeing quality-of-service in highspeed networks, In: proceeding of INFOCOM', 92, (1992), 2016-2025.

[100] Nazarov A., Melikov A., Pavlova E., Aliyeva S., Ponomarenko L., Analyzing an $M/M/N$ Queueing System with Feedback by the Method of Asymptotic Analysis, Cybernetics and Systems Analysis, 57, (2021), 57-65.

[101] Neuts M. F., Matrix-analytic methods in queuing theory, European Journal of Operational Research, 15(1), (1984), 2-12.

[102] Neuts M. F., Matrix-geometric solutions in stochastic models: an algorithmic approach, Courier Corporation,(1994).

[103] Neuts M. F., Matrix geometric solutions in stochastic models: An algorithmic approach, Johns Hopkins University Press, Baltimore, MD, USA, (1981).

[104] Ojobor S. A., Ogini N. O., Threshold recovery policy for the machine interference repair problem with server vacations, Journal of Physics: Conference Series, 2199, (2022),

[105] Raj R., Jain V., Optimization of Traffic Control in $MMAP[k]/PH[k]/S$ Catastrophic Queueing Model with PH Retrial Times and Controllable Preemptive Repeat Policy, arXiv preprint arXiv:2203.02508, (2022).

[106] Panda G., Goswami V., Banik A. D., Equilibrium behaviour and social optimization in Markovian queues with impatient customers and variant of working vacations, RAIRO-Operations Research, 51(3), (2017), 685–707.

[107] Pang G., Whitt W., Infinite-server queues with batch arrivals and dependent service times, Probability in the Engineering & Informational Sciences, 26(2), (2012), 197–220.

[108] Park H.M., Yang W.S.,Chae K.C., Analysis of the $GI/Geo/1$ queue with disasters, Stochastic Analysis & Applications, 28(1), (2009), 44–53.

[109] Pekoz E. A., Joglekar N., Poisson traffic flow in a general feedback, J. of Applied Probability, 39, (2002), 630-636.

[110] Pender J., Poisson and Gaussian approximations for multi–server queues with batch arrivals and batch abandonment, Technical report, Cornell University, Ithaca, NY, (2013).

[111] Prabhu J., Hlynka M., A queueing model with servers disguised as customers, International Journal of Statistics and Probability, 13(2), (2024), 55–65.

[112] Priya K. B., Kumar M. S., Geetha M., Prabhu J., Maheshwari V., Rajendran S., Rajendran P., Queueing network model with jockeying to reduce the waiting time in the airport, International Journal of System of Systems Engineering, 11(3-4), (2021), 363-379.

[113] Ravid R., A new look on the shortest queue system with jockeying, Probability in the Engineering and Informational Sciences, 35(3), (2021), 557-564.

[114] Ren J., Breuer L., Stanford D. A., Yu K., Perturbed risk processes analyzed as fluid flows, Stochastic Models, 25(3), (2009), 522-544.

[115] Rupert G., Miller Jr., A contribution to the theory of bulk queues, Journal of the Royal Statistical Society, Series B (Methodological), 21(2), (1959), 320–3370.

[116] Sahakyan V., Vardanyan A., The queue state for multiprocessor system with waiting time restriction, In 2019 Computer Science and Information Technologies (CSIT),(2019), 116-119.

[117] Sahana P., Baburaj C., 'c Policy'Batch Service Queueing System under Bernoulli Scheduled Vacation and Manual Service:'c Policy'Batch Service Queueing System, Journal of the Kerala Statistical Association, 33(1), (2023), 71-87.

[118] Seeniraj G., Jayaraman R., Moganraj, D., Steady state analysis of queueing system with random vacation subject to supplementary variable method, J. Math. Comput. Sci., 11(2), (2021), 1838-1848.

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

[120] Shanbhag D.N., On infinite server queues with batch arrivals, Journal of Applied Probability, 3(1), (1966), 274–279.

[121] Shekhar C., Raina A. A., Kumar A. , Iqbal J., A survey on queues in machining system progress from 2010 to 2017, Yugoslav Journal of Operations Research, 27(4),(2017),391–413. https://doi.org/10.2298/YJOR161117006R

[122] Shin Y.W., Multi-server retrial queue with negative customers and disasters, Queueing Systems, 55(4), (2007), 223-237.

[123] Shoukry E. M., Salwa M. A., Boshra A. S., Matrix geometric method for $M/M/1$ queueing model with and without breakdown ATM machines, American Journal of Engineering Research, 7(1), (2018), 246–252.

[124] Soodan, Singh B., Kumar R., A single server queuing system with correlated reneging and feedback of served customers, Communications in Statistics-Theory and Methods, 51(18), (2022), 6461-6475.

[125] Stewart W. J., Introduction to the numerical solution of Markov chains–princeton University Press, (1994).

[126] Sudhesh R., Transient analysis of a queue with system disasters and customer impatience, Queueing Syst, 66(1), (2010), 95–105.

[127] Sudhesh R., Azhagappan A., Transient analysis of an $M/M/1$ queue with variant impatient behavior and working vacations, Opsearch, 55(3), (2018), 787-806.

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

[129] Sudhesh R., Azhagappan A., Dharmaraja S., Transient analysis of $M/M/1$ queue with working vacation, heterogeneous service and customers' impatience, RAIRO-Oper. Res., 51(3), (2017), 591-606.

[130] Sudhesh R., Priya R.S., An analysis of discrete-time $Geo/Geo/1$ queue with feedback, repair and disaster, International Journal of Operational Research, 35(1), (2019), 37–53.

[131] Sun W., Li S., Guo E. C., Equilibrium and optimal balking strategies of customers in Markovian queues with multiple vacations and N-policy, Applied Mathematical Modelling, 40(1), (2016), 284–301.

[132] Suranga M. I. G., Kalidass K., Jicheng L., Transient analysis of an $M/M/1$ queueing system subjected to multiple differentiated vacations, impatient customers and a waiting server with application to IEEE 802.16E power saving mechanism, Indian145-146 J. Pure Appl. Math., 51(1), (2020), 297-320.

[133] Takács L., A queuing model with feedback, RAIRO-Operations Research-Recherche Oprationnelle, 11(4), (1977), 345-354.

[134] Takács L., A single-server queue with feedback, Bell System Technical J.,42(2), (1963), 505-519.

[135] Takagi H., Queueing Analysis, A Foundation of Performance Evaluation: Vacation and Priority Systems, Volume 1. Elsevier, Amsterdam, (1991).

[136] Takagi H., Takahashi Y. Priority queues with batch Poisson arrivals, Operations Research Letters, 10(4), (1991), 225–232.

[137] Tang S., Modelling and performance analysis of a cloud computing system using an open queueing network with multi-server queues, International Journal of Computer Applications in Technology, 72(1), (2023), 1-12.

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

[139] Tian N., Zhang Z. G., Vacation queueing models, Theory and applications, New York: Springer (2006).

[140] Tikhonenko O., Ziółkowski M., Single–server queueing system with limited queue, random volume customers and unlimited sectorized memory buffer, Bulletin of the Polish Academy of Sciences, Technical Sciences, 70(6), (2022).

[141] Tikhonenko O., Ziéolkowski M., Kempa W. M., Queueing systems with random volume customers and a sectorized unlimited memory buffer, International Journal of Applied Mathematics and Computer Science, 31(3), (2021), 471-486.

[142] Tzenova E. I. , Adan I. J., Kulkarni V. G., Fluid models with jumps, Stochastic Models, 21, (2005), 37-55.

[143] Upadhyaya S., Performance prediction of a discrete-time batch arrival retrial queue with Bernoulli feedback, Applied Mathematics and Computations, 283, (2016), 108–119.

[144] Vadivukarasi M. , Kalidass K., Discussion on the transient behavior of single server Markovian multiple variant vacation queues, Operations Research and Decisions, Wroclaw University of Science and Technology, Faculty of Management, 31(1), (2021), 123-146.

[145] Veatch M. H. , Senning J. R., Fluid analysis of an input control problem, Queueing Systems, 61(2-3), (2009), 87-112.

[146] Vijayashree K. V., Anjuka A., Fluid queue driven by an $M/M/1$ queue subject to Bernoulli-schedule-controlled vacation and vacation interruption, Advances in Operations Research, 2016,(2016), 11.

[147] Vijayashree K. V., Anjuka A., Stationary analysis of an $M/M/1$ driven fluid queue subject to catastrophes and subsequent repair, IAENG International Journal of Applied Mathematics, 43(4), (2013), 238-241.

[148] Wang J., An $M/G/1$ queue with second optional service and server breakdowns, Computers Mathematics with Applications, 47(10-11), (2004), 1713-1723.

[149] Wang K.H., Chang Y.C., Cost analysis of a finite $M/M/R$ queueing system with balking, reneging, and server breakdowns, Math. Methods Op. Res, 56, (2002), 169–180.

[150] Wang K., Li N., Jiang Z., Queueing system with impatient customers: A review. In : Proceedings of 2010 IEEE international conference on service operations and logistics, and informatics, (2010), 82-87.

[151] Wang K. H., Wang T. Y., Pearn W. L., Optimal control of the N policy $M/G/1$ queueing system with server breakdowns and general startup times, Applied Mathematical Modelling, 31(10), (2007), 2199-2212.

[152] Yan K., Kulkarni V., Optimal inventory policies under stochastic production and demand rates, Stochastic Models, 24(2), (2008), 173-190.

[153] Yang D. Y., Chiang Y. C., Tsou C. S., Cost analysis of a finite capacity queue with server breakdowns and threshold-based recovery policy, Journal of Manufacturing Systems, 32(1), (2013), 174-179.

[154] Yang D. Y., Chung C. H., Wu C. H., Sojourn times in a Markovian queue with working breakdowns and delayed working vacations, Comput. Ind. Eng, 156 (13), (2021), 107239.

[155] Ye Q., Liu L., Performance analysis of the $GI/M/1$ queue with single working vacation and vacations, Methods Comp. Appl. Prob, 19, (2017), 685–714.

[156] Yechiali U.,Queues with system disasters and impatient customers when system is down, Queueing Syst,56, (2007), 195–202.

[157] Yen T. C., Wang K. H., Chen J. Y., Optimization Analysis of the N Policy $M/G/1$ Queue with Working Breakdowns, Symmetry, 12 (4),(2020), 583.

[158] Zhang M., Gao S., Queue with working breakdowns and impatient customers, RAIRO-Oper. Res, 54, (2020), 815–825.

[159] Zhang M., Hou Z., $M/G/1$ queue with single working vacation, J. Appl. Math. Comput, 39(1-2), (2012), 221-234.

[160] Zhang Y., Yue D., Sun L., Zuo J., Analysis of the queueing-inventory system with impatient customers and mixed sales, Discrete Dynamics in Nature and Society, (2022).

[161] Zhao Y. Q., The kernel method tail asymptotics analytic approach for stationary probabilities of two-dimensional queueing systems, Queueing Systems, 100(1-2), (2022), 95-131.

[162] Zirem D., Boualem M., Adel-Aissanou K., Aïssani D., Analysis of a single server batch arrival unreliable queue with balking and general retrial time, Quality Technology Quantitative Management, 16(6), (2019), 672-6

ملخص

هذه الأطروحة مخصصة للدراسة المتعمقة لمختلف أنظمة الانتظار الماركوفية متعددة الخوادم المعرضة للخطأ، مما يعكس الديناميكيات المعقدة الكامنة في عمليات الاتصالات الحديثة والأنظمة الصناعية.تتمحور الورقة حول ثلاث دراسات مترابطة: أولاً، ننظر إلى تحليل نظام الانتظار ذو السعة اللانهائية، والذي يتميز بوصول الدفعات، وردود الفعل برنولي، وانهيار النظام، ونفاد صبر العملاء (الذي يتجلى في الرفض والتراجع). نحن نحدد شروط الاستقرار ونستمد حلول الحالة المستقرة باستخدام وظائف توليد الاحتمالات. ويخضع هذا النموذج لتقييم مقاييس الأداء، فضلا عن تحليل مفصل للتكلفة والعائد. تركز الدراسة الثانية على نظام الانتظار ذو السعة المحدودة، ودمج حالات اعطال المجدولة، والاسترداد القائم على العتبة، والعطلات النشطة، وانقطاع عطلة برنولي القانونية، ونفاد صبر العملاء، والاحتفاظ بالعملاء غير الصبر. يتم إجراء تحليل الحالة المستقرة باستخدام طريقة مصفوفة Q ، حيث نقوم باستخلاص مؤشرات الأداء الرئيسية وتحسين دالة التكلفة من خلال البحث المباشر وأساليب شبه نيوتن. يتيح لنا هذا الأسلوب تحديد العدد الأمثل للخوادم، والسعة المثلى للنظام، ومعدلات الخدمة المثلى أثناء فترات الانقطاع والتشغيل العادي. أخيرًا، يركز الجزء الثالث من عملنا على تحليل نموذج قائمة الانتظار M/M/c/N مع ردود فعل برنولي والكوارث والإصلاحات. نحن نقدم حلولاً للحالة العابرة و الحالة المستقرة، مُكمّلةً بتحليل سلس لقائمة الانتظار السوائل. تهدف الأساليب المنهجية المستخدمة في هذه الأطروحة إلى سد الفجوة بين نماذج الطابور النظرية وتطبيقاتها العملية. ومن ثم فإنها تمهد الطريق لتصميم وإدارة أكثر كفاءة للأنظمة في مختلف المجالات، مثل الاتصالات السلكية واللاسلكية والتصنيع وصناعات الخدمات.

الكلمات المفتاحية : أنظمة الطوابير، الأعطال، عدم صبر العملاء، نموذج التكلفة

Abstract

This thesis is dedicated to the comprehensive study of various multi-server Markovian queueing systems subject to breakdowns, thus reflecting the complex dynamics inherent in the operations of modern communication and industrial systems. The document is structured around three interconnected studies: Initially, we focus on analyzing an infinite-capacity queueing system characterized by batch arrivals, Bernoulli feedback, working breakdowns, and customer impatience (disaster through balking and reneging). We establish stability conditions and derive steady-state solutions using probability generating functions. This model undergoes an evaluation of performance measures, as well as a detailed cost-benefit analysis. The second study examines a finite-capacity queueing system, incorporating server breakdowns, threshold-based recovery, working vacations, Bernoulli-schedule vacation interruption, customer impatience, and retention of impatient customers. The steady-state analysis is conducted using the Q matrix method. We derive key performance indicators and optimize the cost function through direct search and quasi-Newton methods. This approach allows us to determine the optimal number of servers, the optimal system capacity, as well as the optimal service rates during breakdown and normal operating periods. Finally, the third part of our work focuses on analyzing an M/M/c/N queueing model with Bernoulli feedback, catastrophes, and repairs. We provide both transient and steady-state solutions, complemented by a fluid queue analysis. The methodological approaches deployed in this thesis aim to bridge the gap between theoretical queueing models and their practical applications. They thus pave the way for more efficient system design and management in various fields, such as telecommunications, manufacturing, and service industries.

Keywords: Queueing systems, breakdowns, customers' impatience, cost model.

Résumé

Cette thèse se consacre à l'étude approfondie de divers systèmes de files d'attente Markoviens multi-serveurs sujets à des pannes, reflétant ainsi la dynamique complexe inhérente aux opérations des systèmes de communication et industriels modernes. Le document s'articule autour de trois études interconnectées: Dans un premier temps, nous nous penchons sur l'analyse d'un système de files d'attente à capacité infinie, caractérisé par des arrivées par lots, Bernoulli feedback, des pannes de travail et l'impatience des clients (manifestée par le balking et le reneging). Nous établissons les conditions de stabilité et dérivons les solutions à l'état stationnaire à l'aide de fonctions génératrices de probabilités. Ce modèle fait l'objet d'une évaluation des mesures de performance, ainsi que d'une analyse coût-bénéfice détaillée. La deuxième étude porte sur un système de files d'attente à capacité finie, intégrant des pannes de serveur, une récupération basée sur un seuil, des vacances actives, une interruption des vacances selon la loi de Bernoulli, l'impatience des clients et la rétention des clients impatients. L'analyse à l'état stationnaire est effectuée en utilisant la méthode de la matrice Q. Nous dérivons des indicateurs de performance clés et optimisons la fonction de coût par le biais de méthodes de recherche directe et quasi-Newton. Cette approche nous permet de déterminer le nombre optimal de serveurs, la capacité optimale du système, ainsi que les taux de service optimaux pendant les périodes de pannes et de fonctionnement normal. Enfin, la troisième partie de notre travail se concentre sur l'analyse d'un modèle de files d'attente M/M/c/N avec Bernoulli feedback, catastrophes et réparations. Nous proposons des solutions tant transitoires que stationnaires, complétées par une analyse de files d'attente fluides. Les approches méthodologiques déployées dans cette thèse visent à combler l'écart entre les modèles de files d'attente théoriques et leurs applications pratiques. Elles ouvrent ainsi la voie à une conception et une gestion plus efficaces des systèmes dans divers domaines, tels que les télécommunications, la fabrication et les industries de services.

Mots clés: Systèmes de files d'attente, pannes, impatience des clients, modèle de coût.