

Transient and Comparative Analysis of a Single-Server Queueing System with Preventive Maintenance, Bernoulli Feedback and Randomized Vacation Strategy

Wenlong CHEN

School of Mathematics and Statistics, Sichuan University of Science and Engineering, Zigong 643000, China

E-mail: henenong@163.com

Shaojun LAN*

School of Mathematics and Statistics, Sichuan University of Science and Engineering, Zigong 643000, China

E-mail: sjlan89@163.com

Abstract This paper investigates an $M/G/1$ queueing system that integrates preventive maintenance, randomized vacation policy, and Bernoulli feedback mechanism. By applying the law of total probability decomposition and Laplace transform (LT) technique, we derive explicit expressions for the LT of the transient queue length distribution starting from arbitrary initial states. Based on the transient results, the steady-state queue length distribution and the corresponding additional queue length distribution are also obtained. Furthermore, we show that the proposed model reduces to several classical systems under specific parameter settings. Finally, numerical experiments are conducted to compare the proposed model with classical vacation models, to examine the system capacity optimization design under loss constraints, and to illustrate the transient behavior of the system. These findings verify the validity and practical relevance of the considered model.

Keywords $M/G/1$ queueing system; randomized vacation strategy; preventive maintenance; stochastic decomposition property; Bernoulli feedback

1 Introduction

Queueing phenomena are prevalent in modern service systems and appear across diverse aspects of daily life. This ubiquity has established queueing theory as a fundamental analytical framework for evaluating service efficiency and optimizing energy and resource allocation. The theory has been extended to model a wide range of stochastic systems in areas including computer networks, manufacturing, transportation, healthcare, and cloud computing. In real-world

Received October 9, 2025, accepted December 29, 2025

Supported by Program of Graduate Education and Teaching Reform in Sichuan University of Science and Engineering (JG202439)

*Corresponding author

operations, servers often suspend service temporarily after completing a busy period, whether for maintenance, personnel breaks, or resource reallocation. This operational pattern is captured by the concept of server vacations, formalized through vacation queueing policies. These policies allow for more realistic modeling of systems with intermittent service, further enhancing the broad applicability of queueing theory in complex, dynamic environments. Levy and Yechiali^[1] pioneered the incorporation of server vacations into queueing theory by extending the classical $M/G/1$ model, providing a systematic analysis of key performance measures. They formally introduced and analyzed both single and multiple vacation policies, establishing a foundational framework that has since inspired extensive research in vacation-based queueing systems and stochastic service control.

Building on the foundational work in vacation queueing systems, significant theoretical and applied advancements have been made by researchers such as Doshi^[2], Takagi^[3], Upadhyaya^[4], Tian and Zhang^[5], and Ke et al.^[6], among others. Their contributions have greatly expanded the analytical toolkit and enhanced the performance evaluation of queueing models incorporating various vacation policies, including single, multiple, and working vacations. These models have demonstrated broad applicability across diverse domains, such as telecommunication networks, manufacturing systems with setup or repair times, and energy-efficient computing environments, where managing periods of resource idleness is critical for optimizing system efficiency.

As queueing theory continues to evolve, increasing attention has been devoted to accurately modeling server behavior during idle or vacation periods—phases that are crucial for evaluating system performance under realistic operating conditions. In practical systems such as manufacturing lines, telecommunications networks, and computing infrastructures, servers often undergo various forms of inactivity due to operational protocols, energy-saving mechanisms, or control delays. These inactive periods frequently differ in both nature and duration, particularly between the initial idle period following a busy cycle and subsequent idle or vacation intervals. Traditional vacation models, which assume that all vacation durations are identically distributed, are insufficient to capture this behavioral asymmetry. To overcome this limitation, researchers have developed a more flexible theoretical framework by relaxing the assumption of identical distributions. A key advancement in this direction is the introduction of heterogeneous vacation models^[5,7–11], where the first vacation is governed by a different probability distribution than subsequent vacations. This enhanced modeling capability enables a more realistic representation of system dynamics across distinct operational phases, yielding deeper insights into critical performance metrics such as queue length, waiting time, and server utilization. Tian and Zhang^[5] conducted a systematic study on heterogeneous vacation structures and revealed that the distributional differences between the first and subsequent vacations significantly affected steady-state behavior and key performance measures. Li and Zhu^[7] introduced a delayed vacation mechanism based on the classical $M/G/1$ queue with vacations, in which the server, upon completion of a busy period, enters an idle waiting phase. If no new arrivals occur during this phase, the server proceeds to a vacation following a predefined mixed policy. They

derived the LT expressions of the queue length distribution and the residual service time at arbitrary time points. Hassan and Atiquzzaman^[8] further extended this model by incorporating a setup time for service resumption, enabling joint modeling of delayed vacations and server startup processes. Following this line of research, Kumar et al.^[9] proposed an $M/G/1$ model incorporating setup time, shutdown time, and a vacation continuation mechanism governed by probabilistic rules. Using an embedded Markov chain approach, they derived exact expressions for the probability generating function of steady-state distribution of the system and evaluated performance metrics such as the average number of customers per service cycle. Recent work Vignesh et al.^[10] further extended the modeling of service systems by incorporating startup delays, thereby enhancing their flexibility and realism in representing practical operational dynamics.

Beyond structural enhancements such as setup times and heterogeneous vacation durations, a distinct research stream has focused on the strategic uncertainty in server behavior at the end of a vacation. In classical vacation queueing models, such as the single-vacation and multiple-vacation policies, the server's post-vacation action is deterministic: in the single-vacation model, the server takes exactly one vacation after each busy period and then returns to serve waiting customers; in the multiple-vacation model, the server continues taking successive, identically distributed vacations until it finds customers upon return. In contrast, more recent models introduce adaptive decision-making at vacation completion. Rather than following a fixed rule, the server may either resume service or initiate another vacation based on probabilistic or state-dependent criteria. A widely adopted framework is the Bernoulli vacation policy, where, upon returning from a vacation, the server resumes service with probability p and begins another vacation with probability $1 - p$. This approach not only subsumes the classical single- and multiple-vacation models as limiting cases (when $p = 1$ and $p = 0$, respectively), but also enables a continuous spectrum of server behaviors in between. Compared to fixed-vacation models, the Bernoulli rule offers greater modeling flexibility and realism, particularly in systems where service resumption is uncertain or deliberately delayed, such as energy-saving modes in internet of things (IoT) devices, delayed wake-up policies in cloud servers, or human-in-the-loop service systems with probabilistic availability. It also facilitates fine-grained control over trade-offs between energy efficiency and response time, making it a powerful tool in the design of adaptive service systems. Choudhury and Deka^[12] analyzed an $M^X/G/1$ repairable queueing model with two phases of service and Bernoulli vacation schedule under multiple vacation policy. They derived queue size distributions at various epochs, busy period distribution, waiting time distribution under the steady-state condition. Gao and Liu^[13] proposed an $M/G/1$ queue with Bernoulli interruption, where the continuation or termination of a task depends on probabilistic rules, offering greater flexibility in system dynamics. Jain and Kaur^[14] focused on an unreliable server retrial queueing model with Bernoulli vacation, balking customers, and optional services. They evaluated queue size distribution and performance metrics using the supplementary variable technique (SVT), and employed the maximum entropy principle (MEP) for approximation analysis. In addition to the above studies, a number of recent works have

examined queueing systems that incorporate mechanisms such as server vacations, service interruptions, retrial behavior, and reliability-oriented controls. Xu et al.^[15] analyzed an $M/G/1$ queue with a working-vacation service policy. Seeniraj et al.^[16] studied steady-state behavior in a service-interruption framework. Baskar and Saravanarajan^[17] investigated a double-orbit retrial model with working vacations. Nascimento et al.^[18] addressed probabilistic service rules in an $M/G/1$ setting. Houari et al.^[19] considered unreliable servers combined with vacation schedules. Keerthana^[20] analyzed a multi-attribute $M/G/1$ queue with interruptions. Liu et al.^[21] examined transient behavior in a feedback-based vacation model. These works collectively reflect the recent trend of enriching classical queueing models with operational features such as vacations, interruptions, feedback, and adaptive service mechanisms.

While classical vacation models provide a valuable theoretical foundation for queueing analysis, they often rely on oversimplified assumptions about server behavior. Specifically, they assume that the server immediately enters a vacation upon system idleness and resumes service deterministically at the end of the vacation period. These assumptions overlook critical operational realities, particularly the strategic use of idle periods for preventive maintenance (PM)—a practice essential for ensuring long-term system reliability, minimizing unexpected failures, and reducing costly downtime.

Preventive maintenance plays a vital role in numerous real-world systems. For example, in manufacturing systems, computer numerical control (CNC) machines undergo routine lubrication, calibration, and component inspections during idle times to maintain precision and avoid sudden breakdowns that could halt production lines. In cloud computing data centers, servers are taken offline during low-traffic periods to perform firmware updates, disk defragmentation, and thermal checks, thereby enhancing energy efficiency and system longevity. Similarly, in railway signaling systems and air traffic control, maintenance windows are scheduled during off-peak hours to ensure safety-critical equipment remains fully operational during peak demand.

By neglecting preventive maintenance, classical queueing models fail to capture the trade-offs between service availability, system reliability, and resource utilization, especially in safety-critical or resource-constrained environments. To bridge this gap, recent studies have integrated preventive maintenance into queueing frameworks, allowing servers to perform reliability-enhancing tasks during idle or vacation periods. This advancement not only improves model realism but also supports the design of more resilient and sustainable service systems. Hsu^[22] explored optimal maintenance strategies in an unreliable $M/G/1$ production system using renewal theory. Co and Araar^[23] investigated the trade-off between maintenance frequency and system reliability, showing that shortening the interval between maintenance activities can effectively reduce the incidence of system failures. However, excessive maintenance may lead to higher operational costs, highlighting the need for an optimized maintenance scheduling strategy. They established a theorem for identifying the maintenance interval that minimizes cost in a specific setting. Recently, Rathinasamy et al.^[24] analyzed a retrial queueing system subject to random service interruptions and preventive maintenance. They presented numerical

evaluations using the supplementary variable method and confirmed that the system exhibited stochastic decomposition behavior. Wang et al.^[25] conducted the cost-benefit analysis of four retrial systems that incorporate preventive maintenance and unreliable service stations, which aims to assess the performance and efficiency of these systems under varying conditions. Their findings indicated significant differences in the reliability and availability of the retrial systems based on their configurations.

Existing vacation queueing models have long served as foundational tools in the performance analysis of stochastic service systems, with broad applications in manufacturing, telecommunications, and energy-aware computing. However, many of these models rely on idealized assumptions, such as deterministic vacation scheduling and perfect server reliability, which limit their fidelity in dynamic, real-world environments. In particular, they often fail to account for critical operational practices such as preventive maintenance, customer feedback due to dissatisfied or incomplete service, and adaptive server idleness control, which collectively influence system reliability, efficiency, and responsiveness.

To address these limitations, this paper proposes a novel $M/G/1$ queueing model with integrated randomized vacation strategy, Bernoulli feedback, and scheduled preventive maintenance—a combination that, to the best of our knowledge, has not been studied in a unified analytical framework. The proposed model features a randomized vacation policy in which the server probabilistically decides to resume service or initiate another vacation upon return, thereby generalizing classical single- and multiple-vacation models. Concurrently, periodic preventive maintenance is scheduled during idle periods to enhance system reliability and mitigate unexpected failures, while Bernoulli feedback captures the behavior of customers who, after unsatisfactory service, re-enter the system with a certain probability.

The proposed model offers a realistic representation of automated service management in modern cloud data centers. In such systems, each server is responsible for processing client requests for computing or data services. After completing a task, if the result is unsatisfactory or the transmission fails due to transient errors such as internal software failures or brief connectivity issues, the task is automatically returned to the service queue for reprocessing. This kind of repeated request handling is captured by the Bernoulli feedback mechanism in the proposed model, where each customer has a fixed probability of requiring additional service. In addition to processing tasks, servers regularly undergo preventive maintenance to ensure continued reliability and security. These operations include updating system software, running security diagnostics, and performing hardware checks. During maintenance, the server is temporarily unavailable, and this downtime is explicitly modeled through the preventive maintenance component of the queueing framework. Following the completion of maintenance or a service session, the server evaluates whether to stay active in anticipation of new requests or to transition into a low-power state in order to reduce energy consumption. This decision is typically made according to a probabilistic policy influenced by system conditions, including the current arrival rate and recent task history. Such probabilistic behavior aligns with the standby-or-vacation

mechanism described in the model. When the arrival of new tasks is infrequent, a server may repeatedly enter energy-saving sleep modes. In this case, it wakes intermittently to check for new arrivals, and if no requests are detected, it continues to idle across multiple periods until a new task appears. This sequence of repeated inactive states is accurately captured by the multiple vacation structure in the proposed model. By integrating these operational mechanisms, including feedback-driven reprocessing, scheduled preventive maintenance, probabilistic standby decisions, and repeated server vacations, the model provides a comprehensive analytical tool for evaluating performance in high-reliability and energy-aware computing environments. Conventional $M/G/1$ queueing models are generally not capable of simultaneously describing these interacting features. Therefore, the unified modeling approach presented in this study is more suitable for capturing the dynamics of modern service systems.

The main contributions of this study are as follows:

- 1) The proposed queueing system introduces a probabilistic control structure for server operations by combining non-disruptive preventive maintenance, randomized vacation strategy, and Bernoulli feedback. This unified framework extends the existing vacation models and enables greater flexibility in modeling complex real-world service systems.
- 2) This study employs the law of total probability decomposition in conjunction with LT techniques. These tools allow for rigorous derivation of performance metrics under general assumptions without requiring steady-state conditions at the outset.
- 3) Instead of directly focusing on steady-state characteristics, the paper first investigates the transient queue length distribution from arbitrary initial states. This transient analysis provides a foundation for recursively deriving steady-state results, which is often challenging in queueing analysis.
- 4) Numerical studies are carried out to validate the theoretical analysis, optimize the system capacity, and compare the steady-state queue length distribution with classical models. The results confirm the enhanced performance of the proposed system in mitigating congestion and improving service efficiency.

The remainder of this paper is organized as follows. Section 2 introduces the system description and presents some definitions and lemmas. In Section 3, we derive the transient queue length distribution under general initial conditions. Section 4 focuses on the steady-state analysis, in which the expressions for the queue length distribution and its probability generating function are obtained. In Section 5, we discuss several special cases, showing that the proposed model can reduce to classical models under specific parameter settings. Section 6 verifies the correctness of the theoretical analysis and illustrates the practical value of the theoretical results through numerical analysis. Finally, Section 7 concludes the paper and discusses potential directions for future research.

2 Model Description and Preliminaries

This section presents the formulation of the considered queueing model along with the necessary preliminary groundwork. The first part provides a detailed description of the model structure and associated notation. The second part presents the preparatory materials required for subsequent analysis, including key definitions and lemmas.

2.1 Model Description

We consider an $M/G/1$ queue system with preventive maintenance, randomized vacation strategy and Bernoulli feedback. A detailed description of the model is provided below, outlining its key components and operational rules.

1) Customers arrive at the system according to a Poisson process with rate ξ . Let $\{\tau_i, i = 1, 2, \dots\}$ denote the interarrival times, which are independent and identically distributed (i.i.d.) exponential random variables with rate ξ . That is, the cumulative distribution function (CDF) of interarrival times is

$$F(t) = P\{\tau_i \leq t\} = 1 - e^{-\xi t}, \quad t \geq 0, \quad i = 1, 2, \dots \quad (1)$$

2) Customers are served by a single server operating under a first-come, first-served (FCFS) discipline. The service times $\{\chi_i, i = 1, 2, \dots\}$ are i.i.d. random variables with general distribution $G(t) = P\{\chi_i \leq t\}$ and mean $\mu^{-1} = \mathbb{E}[\chi_i] = \int_0^\infty t dG(t) < \infty, i = 1, 2, \dots$.

3) Whenever the system becomes empty, the server performs a non-disruptive preventive maintenance with a random duration Φ , which follows a general distribution $\Phi(t), t \geq 0$. When the preventive maintenance is completed, if there are customers in the system, the server immediately resumes service; otherwise, the server will either enter a standby period $\aleph$ with probability ϖ , during which it remains available to serve arriving customers immediately, or begin to take vacations according to a multiple vacation policy with probability $1 - \varpi$. Specifically, if no customers are present in the system at the end of a vacation, the server continues to take successive vacations until it returns and finds at least one customer waiting, at which point the vacation ends. The vacation time series $\{V_i, i = 1, 2, \dots\}$ consists of independent random variables following a general distribution $V(t), t \geq 0$.

4) Customers, upon completion of service, depart the system forever with a probability ϱ or re-enter the queue to request service again with a probability $1 - \varrho$. That is to say, the Bernoulli feedback mechanism is adopted.

5) To reflect practical scenarios more accurately, we assume that if the system is empty at time $t = 0$, the server does not perform maintenance immediately, but remains in the system to wait for the arrival of the first customer. Furthermore, all random variables associated with the system are mutually independent.

We now introduce the basic notations used throughout the analysis in the Table 1.

2.2 Preliminaries

To facilitate subsequent analysis, we introduce some definitions and lemmas.

Table 1 Summary of the notation used in the model formulation

Symbol	Description
ξ	Rate of Poisson arrival process
$\tau_i, F(t)$	Interarrival time of the i -th customer, and its CDF
$\chi_i, G(t), g(s)$	Service time of the i -th customer, the CDF $G(t)$ of χ_i , and the Laplace–Stieltjes transform(LST) of $G(t)$ defined by $g(s) = \int_0^\infty e^{-st} dG(t)$
μ^{-1}	Mean service time
$\Phi, \Phi(t), \phi(s)$	Preventive maintenance time, the CDF of Φ , and the LST of $\Phi(t)$ defined by $\phi(s) = \int_0^\infty e^{-st} d\Phi(t)$
ϖ	Randomized vacation probability
$V(t), v(s)$	Vacation time distribution, and its LST defined by $v(s) = \int_0^\infty e^{-st} dV(t)$
ϱ	Bernoulli feedback probability
$\hat{\tau}$	System idle period
M	Number of services required under Bernoulli feedback
$\aleph$	Standby period after preventive maintenance
$\Gamma(t), \bar{\Gamma}(t), \Gamma^{(k)}(t), \gamma(s)$	Generalized service time distribution, its complementary CDF $\bar{\Gamma}(t) = 1 - \Gamma(t)$, the k -fold convolution of $\Gamma(t)$, and the LST of $\Gamma(t)$ defined by $\gamma(s) = \int_0^\infty e^{-st} d\Gamma(t)$
$N(t)$	Number of customers in the system at time t
$b, B(t), b(s), \mathbb{E}[b]$	Busy period triggered by a customer, the CDF of b , and the LST of $B(t)$ defined by $b(s) = \int_0^\infty e^{-st} dB(t)$, and expected value of b
$b^{(i)}$	Busy period initiated by i customers
$Q_j(t), q_j^*(s)$	Probability of j customers in the system during a busy period, and its LT $q_j^*(s) = \int_0^\infty e^{-st} Q_j(t) dt$
ρ	Traffic intensity, $\rho = \xi \mathbb{E}[\Gamma]$
ω	Root of $z = \gamma(\xi - \xi z)$ with $0 < \omega < 1$
$L, \mathbb{E}[L]$	Steady-state queue length, and its expected value
$\Re(s)$	Real part of complex variable s

Definition 1 System idle period: The time interval from the moment the system becomes empty until the arrival of the next customer.

Since customer arrivals occur according to a Poisson process with rate ξ , the distribution of the system idle period $\hat{\tau}$ is given by

$$P\{\hat{\tau} \leq t\} = 1 - e^{-\xi t}, \quad t \geq 0. \quad (2)$$

Definition 2 Server idle period: The time interval starting from the moment the system becomes empty until the server begins serving the next customer, which may include preventive maintenance or vacation periods depending on the operational mechanism.

Definition 3 Generalized service time: The total time a customer spends in service from the start of the first service until ultimate departure from the system. Due to the Bernoulli feedback mechanism, after each service completion the customer re-enters the queue with prob-

ability $1 - \rho$ or leaves with probability ρ . Hence, the generalized service time is the sum of the durations of possible multiple services times.

Let M denote the total number of services received by a customer, which follows a geometric distribution with success probability ρ , i.e.,

$$P\{M = k\} = \rho(1 - \rho)^{k-1}, \quad k = 1, 2, \dots. \quad (3)$$

According to the previous model assumptions, the total number of services M is independent of the sequence of service times $\{\chi_i, i = 1, 2, \dots\}$. Then the generalized service time of a customer, denoted by Γ , is expressed as

$$\Gamma = \sum_{i=1}^M \chi_i, \quad i = 1, 2, \dots. \quad (4)$$

Let $\Gamma(t) = P\{\Gamma \leq t\}$ denote the distribution function of Γ . The LST of $\Gamma(t)$, denoted by $\gamma(s)$, is given by

$$\gamma(s) = \frac{\rho g(s)}{1 - (1 - \rho)g(s)}. \quad (5)$$

Definition 4 Server busy period: The time interval that begins when the server initiates service on a customer (leading the server to transition from an idle to a busy state) and ends when the system becomes empty again. The duration of the busy period is fundamentally shaped by the Bernoulli feedback mechanism, since customers who return to the queue create extra service requirements, increasing the server's workload and extending the busy period compared to a system without Bernoulli feedback.

Let b represent the length of the server busy period starting from a customer, with distribution function

$$B(t) = P\{b \leq t\}, \text{LST: } b(s) = \int_0^\infty e^{-st} dB(t). \quad (6)$$

Based on the analytical method proposed in [21], we derive the following lemma 1.

Lemma 1 For $\Re(s) > 0$, $b(s)$ is the unique root of the equation $z = \gamma(s + \xi - \xi z)$ within the region $|z| < 1$. Then the distribution function is

$$B(t) = \sum_{k=1}^{\infty} \int_0^t \frac{(\xi x)^{k-1}}{k!} e^{-\xi x} d\Gamma^{(k)}(x), \quad (7)$$

where $\Gamma^{(k)}(x)$ denotes the k -fold convolution of the distribution function $\Gamma(x)$ for $k = 1, 2, \dots$, with the convention $\Gamma^{(0)}(x) = 1$. Moreover, we have

$$\lim_{t \rightarrow \infty} B(t) = \lim_{s \rightarrow 0^+} b(s) = \begin{cases} 1, & \rho \leq 1, \\ \omega < 1, & \rho > 1, \end{cases} \quad \mathbb{E}[b] = \begin{cases} \frac{\rho}{\xi(1-\rho)}, & \rho < 1, \\ \infty, & \rho \geq 1, \end{cases} \quad (8)$$

where ω ($0 < \omega < 1$) is the root of the equation $z = \gamma(\xi - \xi z)$.

Let $b^{<i>}$ denote the server busy period initiated by i ($i = 1, 2, \dots$) customers. Since customer arrivals follow a Poisson process and are independent of the service times, $b^{<i>}$ can be expressed as $b^{<i>} = \sum_{j=1}^i b_j$, where each b_j follows the same distribution as b . Thus, the probability distribution function of $b^{<i>}$ is given by

$$P\{b^{<i>} \leq t\} = B^{(i)}(t), \quad t \geq 0, \quad i \geq 1. \quad (9)$$

Let $Q_j(t) = P\{b > t \geq 0, N(t) = j\}$ denote the transient probability that the server remains busy at time t during a busy period b , and there are exactly j customers in the system, where $j = 1, 2, \dots$. Assume that at time $t = 0$, the busy period has just begun and there is exactly one customer in the system. That is, the initial conditions are $Q_1(0) = 1, Q_j(0) = 0, j = 2, 3, \dots$.

Using the analytical approach developed in [26], we obtain the following lemma 2.

Lemma 2 *Let $q_j^*(s)$ be the LT of $Q_j(t), j = 1, 2, \dots$. For $\Re(s) > 0$, we have*

$$q_j^*(s) = \frac{b(s) - (1 - \varrho)g(s + \xi)b(s)}{\varrho g(s + \xi)} \int_0^\infty e^{-(s+\xi)t} \bar{\Gamma}(t) \frac{(\xi t)^{j-1}}{(j-1)!} dt + \frac{1 - (1 - \varrho)g(s + \xi)}{\varrho g(s + \xi)} \sum_{k=1}^{j-1} \frac{q_{j-k}^*(s)}{b^k(s)} \left\{ b(s) - \sum_{i=0}^k \int_0^\infty e^{-(s+\xi)t} \frac{[\xi b(s)t]^i}{i!} d\Gamma(t) \right\},$$

where $g(s + \xi) = \int_0^\infty e^{-(s+\xi)t} dG(t)$, and $\sum_{k=i}^j \cdot = 0$, if $j < i$.

3 Transient Distribution of the System Queue Length

Next, we will discuss the transient probability distribution of the system queue length.

Let $p_{ij}(t) = P\{N(t) = j \mid N(0) = i\}$ represent the transient probability that the queue length is j at time t , given that there are i customers in the system at the initial time 0. Its LT expression is defined as

$$p_{ij}^*(s) = \int_0^\infty e^{-st} p_{ij}(t) dt. \quad (10)$$

We now turn to the transient behavior of the queue length process. Theorem 1 addresses the transient behavior of the queue length distribution, specifically the probability that the system is empty at time t (i.e., the queue length equals zero), under any arbitrary initial condition. Theorem 2 analyzes the transient queue length distribution for scenarios where the queue length is greater than zero at time t , starting from an arbitrary initial state.

Theorem 1 *For $\Re(s) > 0$ and $i = 1, 2, \dots$, we have*

$$p_{00}^*(s) = \frac{1 - f(s)}{s} \cdot \left\{ 1 + \frac{f(s)b(s)[1 - v(s + \xi)]}{h(s)} \right\}, \quad (11)$$

$$p_{i0}^*(s) = \frac{1 - f(s)}{s} \cdot \left\{ 1 + \frac{b^i(s)[1 - v(s + \xi)]}{h(s)} \right\}, \quad (12)$$

where

$$h(s) = 1 - v(s + \xi) - [1 - v(s + \xi)] \sum_{n=1}^{\infty} b^n(s) \int_0^\infty e^{-(s+\xi)t} \frac{(\xi t)^n}{n!} d\Phi(t) -$$

$$\begin{aligned} & \varpi b(s)[1 - v(s + \xi)] \int_0^\infty e^{-st} \Phi(t) dF(t) - \\ & (1 - \varpi) \sum_{n=1}^\infty b^n(s) \phi(s + \xi) \int_0^\infty e^{-(s+\xi)t} \frac{(\xi t)^n}{n!} dV(t). \end{aligned}$$

Proof The detailed derivation is provided in Appendix A.

Theorem 2 For $\Re(s) > 0$ and $i, j = 1, 2, \dots$, we have

$$p_{0j}^*(s) = f(s)q_j^*(s) + \frac{f(s)\mathfrak{I}_j(s)}{h(s)}, \quad (13)$$

$$p_{ij}^*(s) = \sum_{k=1}^i b^{k-1}(s)q_{j-i+k}^*(s) + \frac{b^{i-1}(s)\mathfrak{I}_j(s)}{h(s)}, \quad (14)$$

where

$$\begin{aligned} t_j(s) &= [1 - v(s + \xi)]b(s) \int_0^\infty e^{-(s+\xi)t} \bar{\Phi}(t) \frac{(\xi t)^j}{j!} dt + \\ & (1 - \varpi)b(s)\phi(s + \xi) \int_0^\infty e^{-(s+\xi)t} \bar{V}(t) \frac{(\xi t)^j}{j!} dt, \\ \delta_j(s) &= [1 - v(s + \xi)] \sum_{n=1}^\infty \sum_{k=1}^n b^k(s)q_{j-n+k}^*(s) \int_0^\infty e^{-(s+\xi)t} \frac{(\xi t)^n}{n!} d\Phi(t) + \\ & (1 - \varpi) \sum_{n=1}^\infty \sum_{k=1}^n b^k(s)\phi(s + \xi)q_{j-n+k}^*(s) \int_0^\infty e^{-(s+\xi)t} \frac{(\xi t)^n}{n!} dV(t), \\ \mathfrak{I}_j(s) &= \delta_j(s)t_j(s) + \varpi b(s)q_j^*(s)[1 + v(s + \xi)] \int_0^\infty e^{-st} \Phi(t) dF(t). \end{aligned}$$

Proof The detailed derivation is provided in Appendix B.

With the transient results established in Theorems 1 and 2, we are now in a position to derive the steady-state queue length distribution.

4 Steady-state Distribution of the System Queue Length

To understand the steady-state state of the system, we will next discuss the steady-state probability distribution of the system queue length.

Theorem 3 Let $p_j = \lim_{t \rightarrow \infty} P\{N(t) = j\}$ ($j = 0, 1, \dots$) be the steady-state queue length distribution of the system. Then, for any initial state, we have:

1) When $\rho = \frac{\xi}{\varrho\mu} \geq 1$, it follows that $p_j = 0$, $j = 0, 1, \dots$, and thus $\{p_j, j = 0, 1, \dots\}$ does not constitute a probability distribution.

2) When $\rho = \frac{\xi}{\varrho\mu} < 1$, the steady-state distribution p_j ($j = 0, 1, \dots$) satisfies the following recursive expressions

$$p_0 = (1 - \rho) \cdot \frac{1 - v(\xi)}{\Lambda}, \quad (15)$$

$$p_j = \xi(1 - \rho) \cdot \frac{\delta_j + t_j + \varpi q_j(1 - v(\xi))\phi(\xi)}{\Lambda}, \quad j = 1, 2, \dots, \quad (16)$$

where

$$\begin{aligned}
 \Lambda &= \xi[1 - v(\xi)]\mathbb{E}[\Phi] + \xi(1 - \varpi)\phi(\xi)\mathbb{E}[V] + \varpi\phi(\xi)[1 - v(\xi)], \\
 \delta_j &= [1 - v(\xi)] \sum_{k=0}^{j-1} q_{j-k} \left[1 - \sum_{n=0}^k \int_0^\infty e^{-\xi t} \frac{(\xi t)^n}{n!} d\Phi(t) \right] + \\
 &\quad (1 - \varpi)\phi(\xi) \sum_{k=0}^{j-1} q_{j-k} \left[1 - \sum_{n=0}^k \int_0^\infty e^{-\xi t} \frac{(\xi t)^n}{n!} dV(t) \right], \\
 q_j &= \frac{1 - (1 - \varrho)g(\xi)}{\varrho g(\xi)} \int_0^\infty e^{-\xi t} \bar{\Gamma}(t) \frac{(\xi t)^{j-1}}{(j-1)!} dt + \\
 &\quad \frac{1 - (1 - \varrho)g(\xi)}{\varrho g(\xi)} \sum_{k=1}^{j-1} q_{j-k} \left\{ 1 - \sum_{i=0}^k \int_0^\infty e^{-\xi t} \frac{(\xi t)^i}{i!} d\Gamma(t) \right\}, \\
 t_j &= [1 - v(\xi)] \int_0^\infty e^{-\xi t} \bar{\Phi}(t) \frac{(\xi t)^j}{j!} dt + (1 - \varpi)\phi(\xi) \int_0^\infty e^{-\xi t} \bar{V}(t) \frac{(\xi t)^j}{j!} dt.
 \end{aligned}$$

Proof From $p_j = \sum_{i=0}^\infty P\{N(0) = i\} \cdot \lim_{t \rightarrow \infty} P_{ij}(t) = \sum_{i=0}^\infty P\{N(0) = i\} \cdot \lim_{s \rightarrow 0^+} s p_{ij}^*(s)$, utilizing Lemmas 1 and 2, Theorems 1 and 2, and applying L'Hôpital's Rule, the desired result is obtained.

Next, we will derive the probability generating function for the steady-state queue length of the system.

Theorem 4 Let $P(z) = \sum_{j=0}^\infty p_j z^j$ denote the probability generating function of the steady-state queue length distribution of the system, and L represent the queue length of the system in steady state. When $\rho < 1$, we have

$$\begin{aligned}
 P(z) &= \frac{\varrho(1 - \rho)(1 - z)g[\xi(1 - z)]}{[\varrho + (1 - \varrho)z]g(\xi(1 - z)) - z} \times \\
 &\quad \left\{ \frac{[1 - v(\xi)][1 - \phi(\xi(1 - z))]}{(1 - z)\Lambda} + \frac{I(z)}{(1 - z)\Lambda} \right\}.
 \end{aligned} \tag{17}$$

The corresponding expected queue length is given by

$$\begin{aligned}
 \mathbb{E}[L] &= \frac{2(1 - \varrho\rho) + \varrho\xi^2\mathbb{E}[\chi^2]}{2(1 - \rho)} + \\
 &\quad \frac{\xi^2[1 - v(\xi)]\mathbb{E}[\Phi^2] + \xi^2(1 - \varpi)\phi(\xi)\mathbb{E}[V^2]}{\Im},
 \end{aligned} \tag{18}$$

where

$$\begin{aligned}
 I(z) &= (1 - \varpi)\phi(\xi)[1 - v(\xi(1 - z))] + \varpi(1 - z)\phi(\xi)[1 - v(\xi)] \\
 \Im &= 2\{\xi[1 - v(\xi)]\mathbb{E}[\Phi] + \xi(1 - \varpi)\phi(\xi)\mathbb{E}[V] + \varpi\phi(\xi)[1 - v(\xi)]\}.
 \end{aligned}$$

Proof By the definition of the probability generating function, we have

$$P(z) = p_0 + \sum_{j=1}^\infty z^j p_j$$

$$= p_0 + \frac{\xi(1-\varpi)}{\Lambda} \left[\sum_{j=1}^{\infty} z^j \delta_j + \sum_{j=1}^{\infty} z^j t_j + \varpi(1-v(\xi))\phi(\xi) \sum_{j=1}^{\infty} z^j q_j \right], \quad (19)$$

where

$$\sum_{j=1}^{\infty} z^j q_j = \frac{z[1-(1-\varrho)g(\xi)][1-g(\xi(1-z))]}{\varrho\xi[g(\xi(1-z))-z]}, \quad (20)$$

$$\begin{aligned} \sum_{j=1}^{\infty} z^j t_j = & \left\{ \frac{z}{1-z}[1-\phi(\xi)] - \frac{1}{1-z}[\phi(\xi(1-z))-\phi(\xi)] \right\} \frac{[1-v(\xi)]}{\xi} + \\ & (1-\varpi) \frac{\phi(\xi)}{\xi} \left\{ \frac{z}{1-z}[1-v(\xi)] - \frac{1}{1-z}[v(\xi(1-z))-v(\xi)] \right\}, \end{aligned} \quad (21)$$

$$\begin{aligned} \sum_{j=1}^{\infty} z^j \delta_j = & [1-v(\xi)] \frac{z[1-(1-\varrho)g(\xi)][1-g(\xi(1-z))]}{\varrho\xi[g(\xi(1-z))-z]} \frac{[1-\phi(\xi(1-z))]}{1-z} + \\ & (1-\varpi)\phi(\xi) \frac{z[1-(1-\varrho)g(\xi)][1-g(\xi(1-z))]}{\varrho\xi[g(\xi(1-z))-z]} \frac{[1-v(\xi(1-z))]}{1-z}. \end{aligned} \quad (22)$$

By substituting Equations (20), (21) and (22) into Equation (19) and simplifying the resulting expression, we obtain Equation (17). Moreover, applying the definition $\mathbb{E}[L] = \frac{d}{dz}P(z)|_{z=1}$, we can derive Equation (18) through straightforward calculation. Thus, the proof of Theorem 4 is completed.

In many vacation queueing systems, the steady-state queue length satisfies a stochastic decomposition structure, as demonstrated in a broad line of studies^[27–32]. This decomposition states that the steady-state queue length can be expressed as the sum of two independent components, where one component reflects the intrinsic behavior of the underlying $M/G/1$ queue and the other captures the additional congestion created by vacations, preventive maintenance, feedback, or other operational mechanisms. The decomposition property provides a powerful analytical insight because it simplifies the study of complex $M/G/1$ models and clarifies the fundamental structure of systems with enriched server behavior. The following Theorem 5 applies this principle to the proposed model by identifying the distribution of the additional queue length generated by the combined effects of preventive maintenance and randomized vacations.

Theorem 5 *In the queueing model examined in this paper, the total steady-state queue length can be decomposed into two components. The first component corresponds to the queue length in a classical $M/G/1$ system with Bernoulli feedback, while the second accounts for the additional queue length caused by preventive maintenance and server's vacations. Let L_d denote the additional steady-state queue length, whose distribution is given as follows:*

$$P\{L_d = 0\} = \frac{1-v(\xi)}{\Lambda}, \quad (23)$$

$$\begin{aligned} P\{L_d = j\} = & \frac{1-v(\xi) - [1-v(\xi)] \sum_{i=1}^j \int_0^{\infty} \frac{(\xi t)^i}{i!} e^{-\xi t} d\Phi(t)}{\Lambda} - \\ & \frac{(1-\varpi)\phi(\xi) \sum_{i=1}^j \int_0^{\infty} \frac{(\xi t)^i}{i!} e^{-\xi t} dV(t)}{\Lambda}, \quad j = 1, 2, \dots. \end{aligned} \quad (24)$$

Proof According to Equation (17), the steady-state queue length of the system studied in this paper can be decomposed into the sum of two independent components. One component is the probability generating function of a classical $M/G/1$ queue with Bernoulli feedback, and the other represents the additional queue length caused by preventive maintenance and server vacations. In the following, we derive the discrete distribution of the additional queue length L_d . Let

$$P_d(z) = \frac{[1 - v(\xi)][1 - \phi(\xi(1 - z))]}{\{\xi[1 - v(\xi)]\mathbb{E}[\Phi] + \xi(1 - \varpi)\phi(\xi)\mathbb{E}[V] + \varpi\phi(\xi)[1 - v(\xi)]\}(1 - z)} + \frac{(1 - \varpi)\phi(\xi)[1 - v(\xi(1 - z))] + \varpi(1 - z)\phi(\xi)[1 - v(\xi)]}{\{\xi[1 - v(\xi)]\mathbb{E}[\Phi] + \xi(1 - \varpi)\phi(\xi)\mathbb{E}[V] + \varpi\phi(\xi)[1 - v(\xi)]\}(1 - z)}. \quad (25)$$

We define

$$H(z) = [1 - v(\xi)][1 - \phi(\xi(1 - z))] + (1 - \varpi)\phi(\xi)[1 - v(\xi(1 - z))] + \varpi(1 - z)\phi(\xi)[1 - v(\xi)],$$

$$I(z) = \frac{1}{1 - z}.$$

Then, we obtain

$$H^{(0)}(z)|_{z=0} = 1 - v(\xi),$$

$$H^{(j)}(z)|_{z=0} = -[1 - v(\xi)] \int_0^\infty (\xi t)^j e^{-\xi t} d\Phi(t) - (1 - \varpi)\phi(\xi) \int_0^\infty (\xi t)^j e^{-\xi t} dV(t), \quad j = 1, 2, \dots,$$

$$I^{(j)}(z)|_{z=0} = j!, \quad j = 0, 1, 2, \dots$$

Combining

$$P\{L_d = j\} = \frac{1}{j!} \cdot \frac{d^j}{dz^j} [P_d(z)]|_{z=0}, \quad (26)$$

with the Leibniz rule for the j -th derivative of a product

$$[H(z)I(z)]^{(j)} = \sum_{k=0}^j \binom{j}{k} H^{(k)}(z) I^{(j-k)}(z), \quad (27)$$

where $H^{(k)}(z)$ denotes the k -th derivative of $H(z)$ with respect to z , and $\binom{j}{k} = \frac{j!}{k!(j-k)!}$, we have

$$P\{L_d = 0\} = \frac{1 - v(\xi)}{\Lambda},$$

$$P\{L_d = j\} = \frac{1}{j!} \cdot \frac{\sum_{k=0}^j \binom{j}{k} H^{(k)}(z) I^{(j-k)}(z)|_{z=0}}{\Lambda}$$

$$= \frac{1}{j!} \cdot \frac{1}{\Lambda} \left\{ [1 - v(\xi)]j! - j \left([1 - v(\xi)] \int_0^\infty \xi t e^{-\xi t} d\Phi(t) + (1 - \varpi)\phi(\xi) \int_0^\infty \xi t e^{-\xi t} dV(t) \right) (j-1)! - \dots - \left([1 - v(\xi)] \int_0^\infty (\xi t)^j e^{-\xi t} d\Phi(t) + \right. \right.$$

$$\begin{aligned}
& (1 - \varpi)\phi(\xi) \int_0^\infty (\xi t)^j e^{-\xi t} dV(t) \Big) \Big\} \\
&= \frac{1}{\Lambda} \left\{ 1 - v(\xi) - [1 - v(\xi)] \sum_{i=1}^j \int_0^\infty \frac{(\xi t)^i}{i!} e^{-\xi t} d\Phi(t) - \right. \\
& \quad \left. (1 - \varpi)\phi(\xi) \sum_{i=1}^j \int_0^\infty \frac{(\xi t)^i}{i!} e^{-\xi t} dV(t) \right\}.
\end{aligned}$$

Thus, the proof of Theorem 5 is finished.

5 Special Cases

To demonstrate the generality and correctness of the proposed model, we now examine two important special cases, which correspond to classical queueing systems with single and multiple vacations under specific parameter settings. By setting certain parameters to specific values in our analytical results, we obtain the desired results that are consistent with those reported in the literature, thus confirming the validity of our framework.

Special case 1.

When $\varrho = 1$ and $\varpi = 1$, the proposed model reduces to the classical single vacation $M/G/1$ queueing system with exhaustive service. Under this setting, the probability generating function of the steady-state queue length distribution and the expected queue length derived from our general expressions are given by

$$\begin{aligned}
P(z) &= \frac{(1 - \rho)(1 - z)g(\xi(1 - z))}{g(\xi(1 - z)) - z} \cdot \frac{1 - \phi(\xi(1 - z)) + (1 - z)\phi(\xi)}{(1 - z)[\phi(\xi) + \xi\mathbb{E}[\Phi]]}, \\
\mathbb{E}[L] &= \rho + \frac{\xi^2\mathbb{E}[\chi^2]}{2(1 - \rho)} + \frac{\xi^2\mathbb{E}[\Phi^2]}{2(\xi\mathbb{E}[\Phi] + \phi(\xi))}.
\end{aligned}$$

These results match exactly with those of the classical single vacation model discussed by Tang^[26].

Special case 2.

When the preventive maintenance period and subsequent vacations follow the same distribution, i.e., $\Phi(t) = V(t)$, and taking $\varrho = 1$ and $\varpi = 0$, the proposed model reduces to the classical $M/G/1$ queueing system with multiple server vacations. Under this setting, we have

$$\begin{aligned}
P(z) &= \frac{(1 - \rho)(1 - z)g(\xi(1 - z))}{g(\xi(1 - z)) - z} \cdot \frac{1 - v(\xi(1 - z))}{\xi(1 - z)\mathbb{E}[V]}, \\
\mathbb{E}[L] &= \rho + \frac{\xi^2\mathbb{E}[\chi^2]}{2(1 - \rho)} + \frac{\xi\mathbb{E}[V^2]}{2\mathbb{E}[V]}.
\end{aligned}$$

Again, these results match exactly with those of the classical multiple vacation model discussed by Tang^[26], thereby validating the generality of our formulation.

6 Numerical Analysis

To validate the analytical results and demonstrate the practical value of the proposed queueing model, this section presents a series of numerical experiments. Firstly, we compare the

proposed system with classical single and multiple vacation models to demonstrate its broader applicability and generalization. Secondly, we focus on the calculations of performance measures and system capacity optimization design. Thirdly, we observe and investigate the situation of the transient queue length. Throughout the numerical study, we assume that the service time follows the exponential distribution $G(t) = 1 - e^{-\mu t}$, the preventive maintenance time follows the exponential distribution $\Phi(t) = 1 - e^{-\theta t}$, and the vacation time follows the exponential distribution $V(t) = 1 - e^{-\varsigma t}$.

6.1 Comparative and Sensitivity Analysis

This subsection presents a comparative analysis between the model studied in this paper and several classical models. By adjusting the parameters ϖ and ϱ , the proposed model can represent various queueing structures. Specifically, we consider the following three models.

- **Model I (single server vacation):** Taking $\varpi = 1$ and $\varrho = 1$, after completing the preventive maintenance during which no customers arrive, the server enters standby mode and waits for the next customer arrival. Under this configuration, the model reduces to the classical single vacation $M/G/1$ queueing model.
- **Model II (multiple server vacations):** When the duration of the preventive maintenance and the subsequent vacations follow the same distribution, i.e., $\Phi(t) = V(t)$, and we set $\varpi = 0$ and $\varrho = 1$, the model reduces to the classical multiple vacations $M/G/1$ queueing model.
- **Model III (the model studied in this paper):** When the standby probability ϖ takes a value in the open interval $0 < \varpi < 1$, the model exhibits a more general structure that incorporates both preventive maintenance and randomized vacation strategy. This setting extends the classical framework in a natural way.

The parameter values used in the simulation for each model are summarized in Table 2.

Table 2 Parameter settings for the three models

Model	ϖ	ϱ	ξ	μ	θ	ς
Model I	1	1	0.04	3.0	0.06	0.06
Model II	0	1	0.04	3.0	0.06	0.06
Model III	0.5	1	0.04	3.0	0.06	0.06

We conducted MATLAB simulations to compute the steady-state queue length distributions under the above configurations. The corresponding results are presented in Table 3.

The numerical results in Table 3 and Figure 1 demonstrate that the proposed model provides a flexible framework that generalizes and outperforms classical single and multiple vacation strategies by adjusting the parameter ϖ . This tunable parameter enables a more nuanced trade-off between service responsiveness and resource efficiency.

Under the setting $\varpi = 1$, which corresponds to the classical single vacation model, the server immediately resumes service after preventive maintenance. This configuration minimizes

Table 3 Steady-state queue length probabilities p_j under three model configurations

p_j	$\varpi = 1.0$	$\varpi = 0.5$	$\varpi = 0.0$	p_j	$\varpi = 1.0$	$\varpi = 0.5$	$\varpi = 0.0$
p_0	0.4465527	0.4363099	0.4252037	p_{10}	0.0000000	0.0000394	0.0000360
p_1	0.3941596	0.3594717	0.3434864	p_{11}	0.0000000	0.0000394	0.0000360
p_2	0.0940654	0.1209541	0.1397520	p_{12}	0.0000000	0.0000000	0.0000000
p_3	0.0386220	0.0512911	0.0550429	p_{13}	0.0000000	0.0000000	0.0000000
p_4	0.0168214	0.0186477	0.0227092	p_{14}	0.0000000	0.0000000	0.0000000
p_5	0.0055174	0.0080426	0.0086511	p_{15}	0.0000000	0.0000000	0.0000000
p_6	0.0027811	0.0031934	0.0031360	p_{16}	0.0000000	0.0000000	0.0000000
p_7	0.0010766	0.0013404	0.0011535	p_{17}	0.0000000	0.0000000	0.0000000
p_8	0.0003140	0.0004337	0.0006128	p_{18}	0.0000000	0.0000000	0.0000000
p_9	0.0000897	0.0002365	0.0001802	$\mathbb{E}[L]$	0.8205715	0.9050266	0.9564559

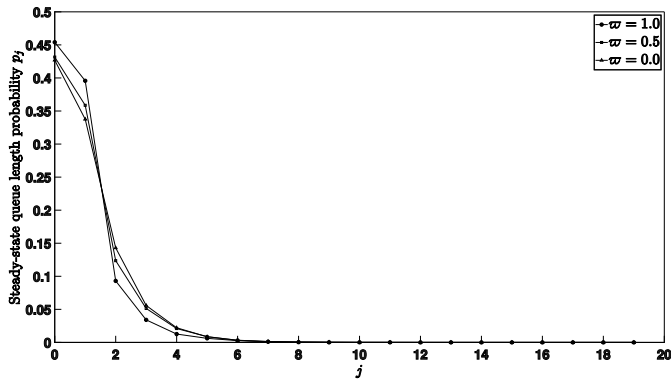


Figure 1 Steady-state queue length distribution p_j under varying standby probabilities ϖ

the average queue length and enhances system responsiveness. However, its main drawback is reduced resource efficiency. The server remains on standby even during extended idle periods, which may lead to unnecessary energy consumption and underutilization of server capacity under low traffic conditions. On the other hand, setting $\varpi = 0$ yields the classical multiple vacations model, in which the server takes a vacation whenever the system becomes idle. If no customers are present when a vacation ends, the server immediately begins a new vacation. This process continues until, at the end of some vacation, the server finds at least one customer in the system and immediately begins serving. This configuration conserves energy by reducing standby time, making it attractive in power-sensitive systems. Nonetheless, this benefit comes at the cost of increased average queue length and slower service response, especially under bursty traffic, which could lead to congestion and degraded quality of service. The intermediate case $\varpi = 0.5$, which defines the general strategy proposed in this study, offers a balanced compromise between these extremes. It achieves a moderate queue length while also allowing controlled waiting behavior. This approach enhances the system’s adaptability to varying traffic intensities and feedback dynamics, ensuring both responsive operation and efficient resource utilization. While classical policies typically prioritize either responsiveness or efficiency, the proposed model incorporates

an adjustable mechanism that balances both objectives. This flexibility makes it especially suitable for stochastic environments with dynamic service demands.

Figure 2 compares the expected queue length $\mathbb{E}[L]$ among the three queueing models as the arrival rate ξ increases. All three curves rise monotonically with ξ , which is consistent with the general behavior of single-server queueing systems under heavier traffic. At the same time, the three models exhibit clearly different growth patterns.

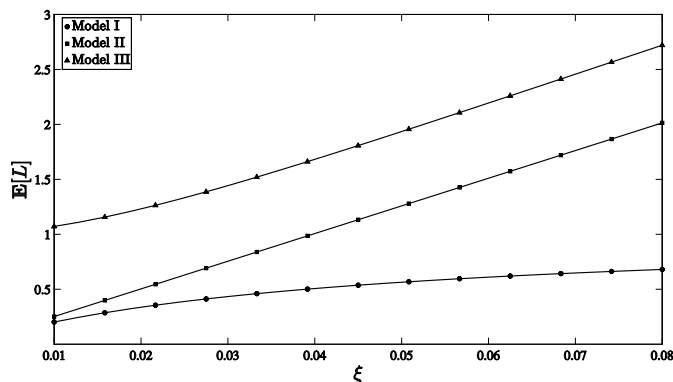


Figure 2 Comparison of $\mathbb{E}[L]$ among three queueing models

Model I produces the smallest values of $\mathbb{E}[L]$, as the server returns to service more promptly. Model II yields intermediate queue length because the repeated vacation mechanism limits the buildup of customers in the system. Model III, which incorporates both preventive maintenance and randomized standby behavior, shows the largest values of $\mathbb{E}[L]$ due to its more flexible and operationally diverse structure. Overall, the comparison highlights how different vacation and standby strategies lead to distinct steady-state queueing behaviors. The proposed Model III provides a generalized framework capable of representing a wider range of system dynamics than the classical single- and multiple-vacation models.

We conducted MATLAB simulations and performed a comparative analysis using ξ and μ as parameters. The corresponding results are shown in Figures 3–5.

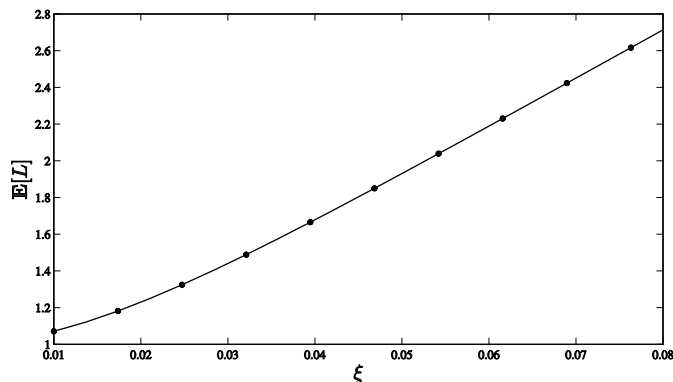


Figure 3 Sensitivity of $\mathbb{E}[L]$ to arrival rate ξ

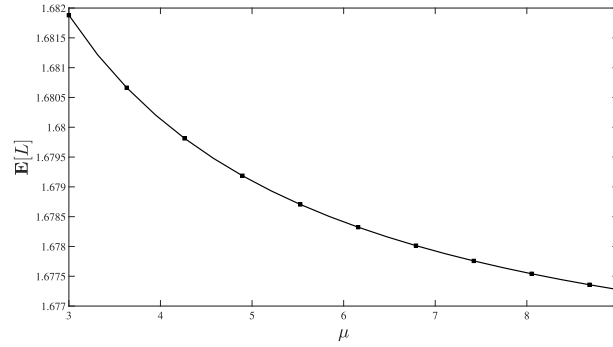


Figure 4 Sensitivity of $\mathbb{E}[L]$ to service rate μ

Figures 3 and 4 illustrate the sensitivity of the expected queue length $\mathbb{E}[L]$ with respect to the arrival rate ξ and the service rate μ . As shown in Figure 3, $\mathbb{E}[L]$ increases almost linearly as ξ grows, indicating that higher arrival intensity leads to increased system congestion. In contrast, Figure 4 shows that $\mathbb{E}[L]$ decreases nonlinearly with increasing μ , reflecting the fact that faster service reduces congestion.

Figure 5 illustrates the variation of steady-state queue length probabilities p_j with respect to the arrival rate ξ .

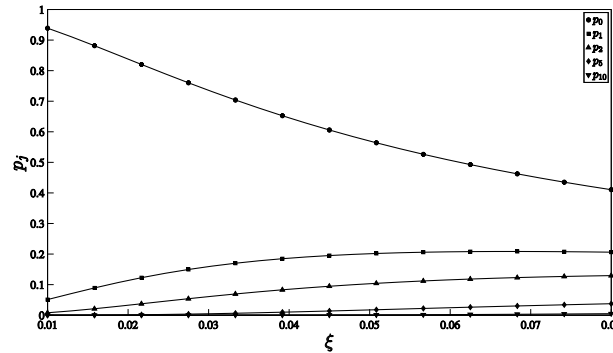


Figure 5 Variation of steady-state probabilities p_j with arrival rate ξ

As shown in the Figure 5, the probability p_0 decreases monotonically as the arrival rate increases. This is consistent with our expected actual situation. Conversely, the probabilities p_j for $j \geq 1$ gradually increase with ξ , indicating that the system is more likely to operate in congested states as customer arrivals become more frequent. Notably, the growth rate of p_j becomes smaller for larger j , and the probability values for higher queue lengths remain relatively low, confirming that the system remains stable within the considered parameter range.

6.2 System Capacity Optimization Design

Given the parameters $\xi = 0.04$, $\mu = 6$, $\theta = 0.04$, $\varsigma = 0.35$, and $\varrho = 0.5$, the steady-state queue length distribution $\{p_j, j = 0, 1, \dots\}$ and the expected queue length $\mathbb{E}[L]$ are computed using MATLAB. The corresponding numerical results are presented in Table 4.

Table 4 Numerical results for steady-state queue length distribution

p_0	p_1	p_2	p_3	p_4	p_5	p_6
0.4563012	0.3383810	0.1071615	0.0476326	0.0237689	0.0115408	0.0069672
p_7	p_8	p_9	p_{10}	p_{11}	p_{12}	p_{13}
0.0040523	0.0021802	0.0007820	0.0005687	0.0003081	0.0002370	0.0000474
p_{14}	p_{15}	p_{16}	p_{17}	p_{18}	p_{19}	$\mathbb{E}[L]$
0.0000474	0.0000237	0.0000000	0.0000000	0.0000000	0.0000000	0.9565856

Using the computed values of p_j , we obtain

$$P\{L > \mathbb{E}[L]\} \approx P\{L \geq 1\} = 1 - p_0 = 1 - 0.4563 = 0.5437, \quad (28)$$

$$P\{L > \mathbb{E}[L] + 2\} \approx P\{L \geq 3\} = 1 - (p_0 + p_1 + p_2) = 0.0981. \quad (29)$$

From Equation (28), it is evident that if the mean queue length is used as a basis for capacity estimation, the customer loss probability reaches 54.37% due to the unavailability of waiting space upon arrival. Even when two additional units are added to the steady-state mean queue length in capacity design, the resulting loss probability, shown in (29), remains as high as 9.81%. This indicates that a significant proportion of customers would still be blocked from entering the system under such a design. Therefore, relying solely on the expected queue length for capacity planning is inadequate and potentially detrimental. To effectively reduce the arrival loss probability and enhance system profitability, a critical question arises: How should the system capacity be dimensioned to achieve a reasonable level?

To meet a exact and specific requirement, e.g., ensuring that the probability of customer loss remains below 0.001, we seek the smallest non-negative integer κ such that $P\{L > \kappa\} \leq 0.001$. According to the numerical results shown in Table 4, this condition is satisfied when $\kappa \geq 10$. Therefore, the system capacity must be designed as 10 to meet basic capacity standard. The above numerical example illustrates that the stationary queue length distribution plays a crucial role in system capacity design. Therefore, deriving an explicit expression for the steady-state queue length distribution $\{p_j, j = 0, 1, \dots\}$ is of significant importance. Furthermore, this example highlights the practical relevance and potential application value of Theorem 3, one of the key contributions of this paper.

To further evaluate the capacity-planning properties of the proposed queueing model, Figure 6 and Table 5 present the minimum system capacity κ required to satisfy the loss-probability constraint $P\{L > \kappa\} \leq 0.001$ for different values of ϖ . All other parameters are fixed at $\mu = 6$, $\theta = 0.04$, $\varsigma = 0.35$, and $\varrho = 0.5$.

Table 5 System optimal capacity for different values of ϖ

ϖ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
κ	9	12	11	12	11	11	9	8	9	9

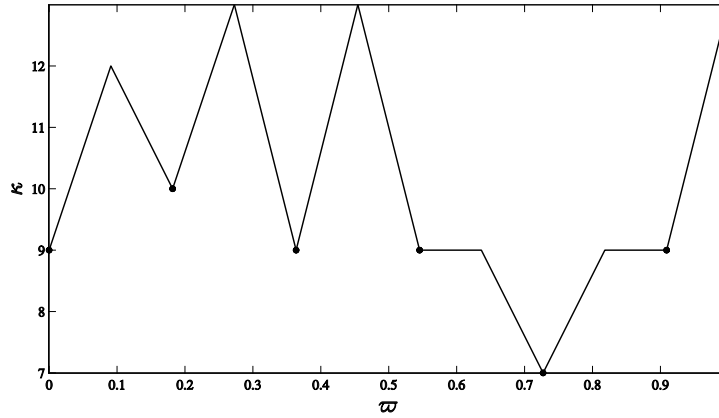


Figure 6 System optimal capacity for different values of ϖ

The results show an overall decreasing tendency in the required capacity κ as the value of ϖ increases, despite some minor fluctuations. A larger ϖ implies that, after preventive maintenance, the server is more likely to remain available instead of entering a vacation period. This improvement in server availability helps limit customer accumulation in the queue, thereby reducing the capacity needed to satisfy the prescribed service-level constraint.

6.3 Transient Behavior Analysis

To analyze the dynamics of the system, we compute the time-dependent state probabilities $P\{N(t) = j\}$ using the LT expressions of $p_{ij}^*(s)$ derived in Theorems 1 and 2. The Gaver–Stehfest numerical inversion algorithm is applied to recover the corresponding time-domain probabilities from their LT expressions (see Kuznetsov^[33] for a detailed convergence analysis and Soundararajan and Barbhuiya^[34] for recent applications in transient queueing analysis). In the numerical experiment, the service time, preventive maintenance time, and vacation time are all assumed to follow exponential distributions with rate parameters $\mu = 2$, $\theta = 0.04$, and $\varsigma = 0.35$, respectively. Customer arrivals are modeled as a Poisson process with rate $\xi = 0.5$, and the randomized vacation and feedback probabilities are set to $\varpi = 0.5$ and $\varrho = 0.5$.

Figure 7 presents the time-dependent evolution of $P\{N(t) = 4\}$ for three different initial states, namely $N(0) = 0$, $N(0) = 5$, and $N(0) = 10$. The results reveal that the transient probabilities differ significantly during the initial phase, reflecting the distinct initial queueing conditions. As time progresses, the trajectories gradually converge toward a common steady-state value, confirming the ergodicity of the system and the validity of the theoretical steady-state analysis presented in Section 4. Figure 7 clearly illustrates how the system evolves from its initial configuration toward equilibrium.

As shown in Figure 7, the transient distribution becomes nearly stable after $t \approx 100$, indicating that the system has reached its steady state. The initial discrepancies among the curves emphasize the dependence of short-term behavior on the initial queue length, which is particularly relevant for systems that frequently restart or operate intermittently.

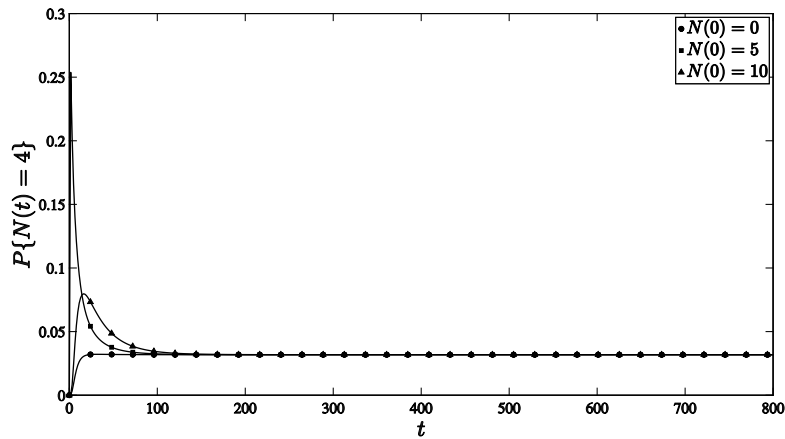


Figure 7 Transient queue length probability $P\{N(t) = 4\}$ for different initial states $N(0) = 0, 5, 10$

Figure 8 presents the transient queue-length probabilities $P\{N(t) = j\}$ for the proposed model with an initial state $N(0) = 4$.

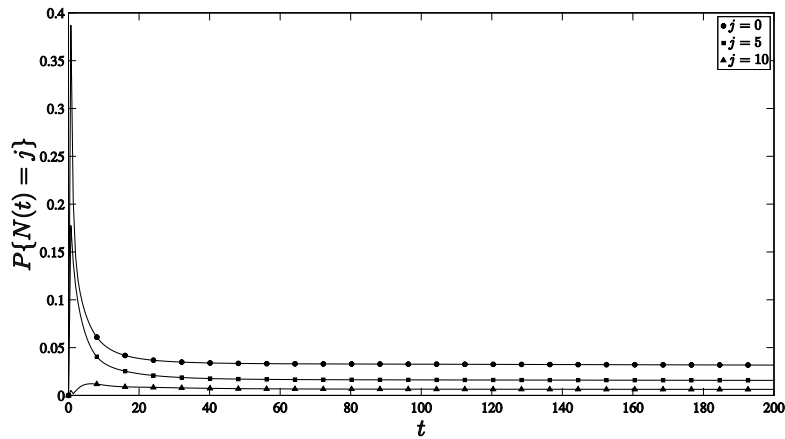


Figure 8 Transient queue length probabilities $P\{N(t) = j\}$ for $j = 0, 5, 10$ under the initial condition $N(0) = 4$

Because the system starts with an initial queue length of $N(0) = 4$, the empty-queue probability $P\{N(t) = 0\}$ remains nearly zero during the initial period, as the system cannot become empty immediately. As time progresses and customers complete service, the probability of entering the empty state gradually increases and eventually converges to its steady-state value. For the moderate queue length $j = 5$, the transient probability exhibits a slight fluctuation at the beginning and then stabilizes, reflecting the system's transition from clearing the initial workload to reaching long-run operation. In contrast, the probability associated with the larger state $j = 10$ remains extremely small throughout the entire time horizon, indicating that such highly congested conditions rarely occur under the current parameter setting. Overall, the results demonstrate that the proposed model exhibits well-behaved transient dynamics and

smoothly converges to its steady-state distribution, with high queue-length states occurring only infrequently.

7 Conclusion

This paper investigates an $M/G/1$ queueing system that incorporates both preventive maintenance and a randomized vacation strategy, together with a Bernoulli feedback mechanism. In this model, once the server completes an uninterrupted preventive maintenance period and finds the system empty, it may either remain in standby mode with a certain probability to wait for incoming customers or begin a sequence of multiple vacations. Additionally, each customer who completes service may re-enter the system based on a geometric feedback mechanism governed by a fixed probability. This setting captures practical scenarios such as rework, inspection loops, or repeated service processes, and offers a more realistic representation than classical single or multiple vacation models, especially in smart manufacturing, IT systems maintenance, and service operations where flexible downtime strategies are essential. Our analysis begins with the study of the transient queue length distribution. By applying the total probability decomposition method and LT techniques, explicit LT expressions for the transient queue length are derived, from which recursive formulas for both the steady-state queue length and the additional queue length are obtained. Several special cases to which the proposed model reduces are also examined. Furthermore, numerical experiments are conducted to illustrate the transient and steady-state behaviors. The transient curves show rapid convergence toward steady-state probabilities and become insensitive to the initial queue length, while the steady-state results generated by the recursive formulas agree well with the numerical computations. A brief sensitivity analysis further reveals that the expected queue length increases with the arrival rate and decreases with the service rate. Finally, a capacity analysis and performance comparison with classical vacation models demonstrate not only the mathematical tractability of the model but also its practical applicability in maintenance-intensive environments such as call centers, data centers, and production systems with feedback loops.

Several directions for future research emerge from this study. One natural extension is to consider more general arrival and service processes, such as Markovian arrival processes (MAP) or phase-type service distributions, to better capture real-world variability. It would also be valuable to investigate customer impatience and abandonment behavior under the proposed model. Furthermore, incorporating dynamic control mechanisms, such as adaptive vacation policies, could enhance system responsiveness and efficiency. Future work may also explore the application of data-driven or learning-based approaches to optimize system performance. We believe that these extensions will not only enrich the theoretical foundation of queueing models, but also improve their practical applicability in areas such as telecommunications, service operations, and intelligent transportation systems.

Acknowledgments

The authors would like to sincerely thank the anonymous reviewers and editors for their valuable comments and constructive suggestions which are very helpful in improving the presentation of this paper.

References

- [1] Levy Y, Yechiali U. Utilization of idle time in an $M/G/1$ queueing system. *Management Science*, 1975, 22(2): 202–211.
- [2] Doshi B T. Queueing systems with vacations: A survey. *Queueing Systems*, 1986, 1: 29–66.
- [3] Takagi H. Queueing analysis: A foundation of performance evaluation. North-Holland, Amsterdam, 1991.
- [4] Upadhyaya S. Queueing systems with vacation: An overview. *International Journal of Mathematics in Operational Research*, 2016, 9(2): 167–213.
- [5] Tian N, Zhang Z G. Vacation queueing models: Theory and applications. Springer Science & Business Media, New York, 2006.
- [6] Ke J C, Wu C H, Zhang Z G. Recent developments in vacation queueing models: A short survey. *International Journal of Operational Research*, 2010, 7(4): 3–8.
- [7] Li H, Zhu Y. Analysis of $M/G/1$ queues with delayed vacations and exhaustive service discipline. *European Journal of Operational Research*, 1996, 92(1): 125–134.
- [8] Hassan M, Atiquzzaman M. A delayed vacation model of an $M/G/1$ queue with setup time and its application to SVCC-based ATM networks. *IEICE Transactions on Communications*, 1997, 80(2): 317–323.
- [9] Kumar B K, Rukmani R, Lakshmi S R A. Performance analysis of an $M/G/1$ queueing system under Bernoulli vacation schedules with server setup and close down periods. *Computers & Industrial Engineering*, 2013, 66(1): 1–9.
- [10] Vignesh P, Srinivasan S, Sundari S M, et al. Study on the queueing characteristics of $M^X/G/1$ queueing model of stages of service with multi vacation policy and setup time phase. *International Journal of Process Management and Benchmarking*, 2023, 15(1): 25–45.
- [11] Ayyappan G, Thilagavathy K. Analysis of MAP/PH(1), PH(2), PH(3)/1 queueing system with two modes of heterogeneous service, standby server, vacation, impatient behaviour of customers, additional service, start-up time, breakdown and phase type repairs. *International Journal of Operational Research*, 2024, 49(1): 85–130.
- [12] Choudhury G, Deka M. A batch arrival unreliable server delaying repair queue with two phases of service and Bernoulli vacation under multiple vacation policy. *Quality Technology & Quantitative Management*, 2018, 15(2): 157–186.
- [13] Gao S, Liu Z. An $M/G/1$ queue with single working vacation and vacation interruption under Bernoulli schedule. *Applied Mathematical Modelling*, 2013, 37(3): 1564–1579.
- [14] Jain M, Kaur S. Bernoulli vacation model for $M^X/G/1$ unreliable server retrial queue with Bernoulli feedback, balking and optional service. *RAIRO - Operations Research*, 2021, 55: S2027–S2053.
- [15] Xu T, Ma Z, Sun Y. Analysis of a repairable $M/G/1$ queue with the gated service and multiple adaptive vacation. *IAENG International Journal of Applied Mathematics*, 2025, 55(5): 1419–1427.
- [16] Seeniraj G, Moganraj D, Jayaraman R. A steady state analysis of single server $M/G/1$ queueing

- system with multiple working vacations. 2nd International Conference on Recent Trends in Applied and Computational Mathematics, 2023, 2649(1): 030026.
- [17] Baskar A, Saravanarajan M C. Analyzing an $M/G/1$ double orbit retrial queue with the implementation of a working vacation policy. *Contemporary Mathematics*, 2024: 3929–3952.
 - [18] Nascimento H J B, Nascimento M E T P, Nascimento D J B, et al. Probabilistic analysis of an $M/G/1$ queue. *Observatório de la Economía Latinoamericana*, 2024, 22(10): e7136–e7136.
 - [19] Houari H, Arrar N, Yahiaoui L. Unreliable $M/G/1$ queue with general retrial time, working vacation and setup time. *Reliability: Theory & Applications*, 2025, 20(1): 543–556.
 - [20] Keerthana R. An $M/G/1$ retrial queue with working vacation, non persistent customers and a waiting server. *Reliability: Theory & Applications*, 2025, 20(1): 1089–1099.
 - [21] Liu Q, Tang Y, Yu M, et al. Analysis of $M^{(\lambda_1, \lambda_2)}/G/1$ queue with uninterrupted single vacation and server's workload controlled D-policy. *Journal of Industrial & Management Optimization*, 2023, 19(6): 4523–4550.
 - [22] Hsu L F. Optimal preventive maintenance policies in an $M/G/1$ queue-like production system. *European Journal of Operational Research*, 1992, 58(1): 112–122.
 - [23] Co H C, Araar A. Age-maintenance of an $M/G/1$ production system. *International Journal of Production Research*, 1991, 29(10): 2135–2149.
 - [24] Rathinasamy P, Krishnasamy K, Karnakumar S, et al. Navigating customer demand in a failure prone retrial queueing system with predictive maintenance. *International Journal of Operational Research*, 2023, 20: 12–24.
 - [25] Wang K H, Wu C H, Yen T C. Comparative cost-benefit analysis of four retrial systems with preventive maintenance and unreliable service station. *Reliability Engineering & System Safety*, 2022, 221: 108342.
 - [26] Tang Y. The transient solution for $M/G/1$ queue with server vacations. *Acta Mathematica Scientia*, 1997, 17(3): 276–282.
 - [27] Fuhrmann S W, Cooper R B. Stochastic decompositions in the $M/G/1$ queue with generalized vacations. *Operations Research*, 1985, 33(5): 1117–1129.
 - [28] Jaya S, Lakshmi B. Stochastic decomposition of the $M/M/1$ queue with environment dependent working vacation. *Journal of Mathematics and Computer Science*, 2021, 11(6): 8008–8020.
 - [29] Kleiner I, Frostig E, Perry D. A decomposition property for an $M^X/G/1$ queue with vacations. *Indagationes Mathematicae*, 2023, 34(5): 973–989.
 - [30] Li L, Liu L, Xu W, et al. Stochastic decomposition in a queueing-inventory system with batch demands, randomized order policy and multiple vacations. *Communications in Statistics – Theory and Methods*, 2024, 53(12): 4407–4427.
 - [31] Majid S, Manoharan P. Analysis of an $M/M/1$ queue with working vacation and vacation interruption. *Applications and Applied Mathematics*, 2019, 14(1): 19–33.
 - [32] Kumar M S, Dadlani A, Moradian M, et al. Age of information in multi-source updating systems: An $M/G/1$ vacation queueing model. *ICC 2023 – IEEE International Conference on Communications*, 2023: 63–68.
 - [33] Kuznetsov A. On the convergence of the Gaver–Stehfest algorithm. *SIAM Journal on Numerical Analysis*, 2013, 51(6): 2984–2998.
 - [34] Soundararajan A, Barbhuiya F P. Transient analysis of a bulk stream queue with arbitrarily distributed arrival intervals. *RAIRO–Operations Research*, 2024, 58(4): 2767–2782.

Appendix

A Proof of Theorem 1

Let $l_0 = 0$, $l_k = \sum_{i=1}^k \tau_i$, $s_0 = 0$, $s_k = \sum_{i=1}^k V_i$, $k = 1, 2, \dots$. It is clear that the system is empty at time t if and only if t lies within the system idle period. By employing the law of total probability decomposition, we obtain

$$\begin{aligned} p_{00}(t) &= P\{0 \leq t < \hat{\tau}_1\} + P\{\hat{\tau}_1 + b_1 \leq t < \hat{\tau}_1 + b_1 + \hat{\tau}_2\} + \\ &\quad P\{t \geq \hat{\tau}_1 + b_1 + \hat{\tau}_2, N(t) = 0\} \\ &= \bar{F}(t) + \int_0^t \bar{F}(t-x) d[F(x) * B(x)] + P\{t \geq \hat{\tau}_1 + b_1 + \hat{\tau}_2, N(t) = 0\}. \end{aligned} \quad (A1)$$

In Equation (A1), the third term represents the system behavior during the server's idle phase following a busy period, which depends on the state of the system at the completion of preventive maintenance and the subsequent server decision. The system can be classified into three cases according to whether a customer arrives during the preventive service period, after the completion of the preventive service when the server immediately starts serving, or during one of the server's vacations following the preventive service.

Case 1: During the preventive maintenance period, customers may arrive, and service will resume after the maintenance is completed. The time t falls within the next system idle period. This situation is illustrated in Figure A1.

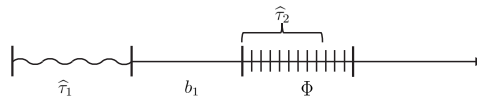


Figure A1 Development process of system under Case 1

Based on this case, the corresponding expression is given by

$$\begin{aligned} &P\{\Phi \geq \hat{\tau}_2, \hat{\tau}_1 + b_1 + \hat{\tau}_2 \leq t, N(t) = 0\} \\ &= \sum_{n=1}^{\infty} P\{\Phi \geq \hat{\tau}_2, \hat{\tau}_1 + b_1 + \hat{\tau}_2 \leq t, 0 \leq l_{n-1} < \hat{\tau}_2 < l_n, N(t-x) = 0\} \\ &= \sum_{n=1}^{\infty} \int_0^t P\{\Phi > \hat{\tau}_2, \hat{\tau}_2 \leq t-x, 0 \leq l_{n-1} < \hat{\tau}_2 < l_n, N(t-x) = 0\} d[F(x) * B(x)] \\ &= \sum_{n=1}^{\infty} \int_0^t \int_0^{t-x} P\{\Phi > y, t-x-y \geq 0, 0 \leq l_{n-1} < y < l_n, N(t-x-y) = 0\} \times \\ &\quad d\Phi(y) d[F(x) * B(x)] \\ &= \sum_{n=1}^{\infty} \int_0^t \int_0^{t-x} p_{n0}(t-x-y) \frac{(\xi y)^n}{n!} e^{-\xi y} d\Phi(y) d[F(x) * B(x)]. \end{aligned} \quad (A2)$$

Case 2: If the system is empty at the end of the preventive maintenance period, the server chooses, with probability ϖ , to remain in the system and wait until a new customer arrives.

Upon the customer's arrival, the server immediately terminates the waiting period and initiates a new busy period. This situation is illustrated in Figure A2.

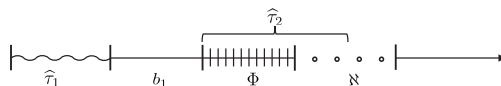


Figure A2 Development process of system under Case 2

Based on this case, the corresponding probability expression is given by

$$\begin{aligned} & \varpi \cdot P\{\hat{\tau}_2 > \Phi, \hat{\tau}_1 + b_1 + \hat{\tau}_2 \leq t, N(t) = 0\} \\ &= \varpi \cdot \int_0^t P\{\hat{\tau}_2 > \Phi, \hat{\tau}_2 < t - x, N(t - x) = 0\} d[F(x) * B(x)] \\ &= \varpi \cdot \int_0^t \int_0^{t-x} p_{10}(t - x - y) \Phi(y) dF(y) d[F(x) * B(x)]. \end{aligned} \quad (\text{A3})$$

Case 3: If the system is empty at the end of the preventive maintenance period, the server chooses, with probability $1 - \varpi$, to take vacations according to a multiple vacation policy. This scenario is shown in Figure A3.

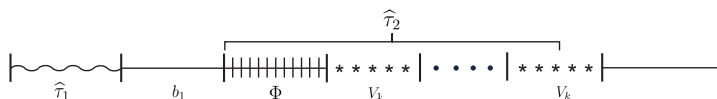


Figure A3 Development process of system under Case 3

Thus, we can get

$$\begin{aligned} & (1 - \varpi) \sum_{k=1}^{\infty} \sum_{n=1}^{\infty} P\{\hat{\tau}_1 + b_1 + \Phi + s_k \leq t, s_{k-1} + \Phi \leq \hat{\tau}_2 < s_k + \Phi, \\ & \quad s_{k-1} + \Phi \leq \hat{\tau}_2 + l_{n-1} < s_k + \Phi \leq \hat{\tau}_2 + l_n, N(t) = 0\} \\ &= (1 - \varpi) \sum_{k=1}^{\infty} \sum_{n=1}^{\infty} \int_0^t P\{\Phi + s_k \leq t - x, s_{k-1} + \Phi \leq \hat{\tau}_2 < s_k + \Phi, \\ & \quad s_{k-1} + \Phi \leq \hat{\tau}_2 + l_{n-1} < s_k + \Phi \leq \hat{\tau}_2 + l_n, N(t - x) = 0\} d[F(x) * B(x)] \\ &= (1 - \varpi) \sum_{k=1}^{\infty} \sum_{n=1}^{\infty} \int_0^t \int_0^{t-x} P\{t - x - y, y \leq \hat{\tau}_2 < y + V_k, \\ & \quad y \leq \hat{\tau}_2 + l_{n-1} < y + V_k < \hat{\tau}_2 + l_n, N(t - x - y) = 0\} \\ & \quad d[\Phi(y) * V^{(k-1)}(y)] d[F(x) * B(x)] \\ &= (1 - \varpi) \sum_{k=1}^{\infty} \sum_{n=1}^{\infty} \int_0^t \int_0^{t-x} \int_0^{t-x-y} P\{t - x - y \geq 0, y \leq \hat{\tau}_2 + l_{n-1} < y + u < \hat{\tau}_2 + l_n, \\ & \quad N(t - x - y - u) = 0\} dV(u) d[\Phi(y) * V^{(k-1)}(y)] d[F(x) * B(x)] \\ &= (1 - \varpi) \sum_{k=1}^{\infty} \sum_{n=1}^{\infty} \int_0^t \int_0^{t-x} \int_0^{t-x-y} p_{n0}(t - x - y - u) \frac{(\xi u)^n}{n!} e^{-\xi(y+u)} dV(u) \times \end{aligned}$$

$$d[\Phi(y) * V^{(k-1)}(y)]d[F(x) * B(x)]. \quad (A4)$$

By substituting Equations (A2), (A3), and (A4) into Equation (A1) and simplifying, we obtain

$$\begin{aligned} p_{00}(t) &= P\{0 \leq t < \hat{\tau}_1\} + P\{\hat{\tau}_1 + b_1 \leq t < \hat{\tau}_1 + b_1 + \hat{\tau}_2\} + P\{t \geq \hat{\tau}_1 + b_1 + \hat{\tau}_2, N(t) = 0\} \\ &= \bar{F}(t) + \int_0^t \bar{F}(t-x)d[F(x) * B(x)] + \\ &\quad \sum_{n=1}^{\infty} \int_0^t \int_0^{t-x} p_{n0}(t-x-y) \frac{(\xi y)^n}{n!} e^{-\xi y} d\Phi(y) d[F(x) * B(x)] + \\ &\quad \varpi \int_0^t \int_0^{t-x} p_{10}(t-x-y) \Phi(y) dF(y) d[F(x) * B(x)] + \\ &\quad (1-\varpi) \sum_{k=1}^{\infty} \sum_{n=1}^{\infty} \int_0^t \int_0^{t-x} \int_0^{t-x-y} p_{n0}(t-x-y-u) \frac{(\xi u)^n}{n!} e^{-\xi(y+u)} dV(u) \times \\ &\quad d[\Phi(y) * V^{(k-1)}(y)]d[F(x) * B(x)]. \end{aligned} \quad (A5)$$

Similarly, for $i = 1, 2, \dots$, we have

$$\begin{aligned} p_{i0}(t) &= \int_0^t \bar{F}(t-x)dB^{(i)}(x) + \sum_{n=1}^{\infty} \int_0^t \int_0^{t-x} p_{n0}(t-x-y) \frac{(\xi y)^n}{n!} e^{-\xi y} dV(y) dB^{(i)}(x) + \\ &\quad \varpi \int_0^t \int_0^{t-x} p_{10}(t-x-y) \Phi(y) dF(y) dB^{(i)}(x) + \\ &\quad (1-\varpi) \sum_{k=1}^{\infty} \sum_{n=1}^{\infty} \int_0^t \int_0^{t-x} \int_0^{t-x-y} p_{n0}(t-x-y-u) \frac{(\xi u)^n}{n!} e^{-\xi(y+u)} \times \\ &\quad dV(u) d[\Phi(y) * V^{k-1}(y)]dB^{(i)}(x). \end{aligned} \quad (A6)$$

Each of these cases contributes to the total response time distribution through different conditional probabilities. Equations (A5) and (A6) describe the system evolution under these conditions. Taking the LT of Equations (A5) and (A6), respectively, we can get

$$\begin{aligned} p_{00}^*(s) &= \frac{[1-f(s)][1+f(s)b(s)]}{s} + f(s)b(s) \sum_{n=1}^{\infty} p_{n0}^* \int_0^{\infty} \frac{(\xi t)^n}{n!} e^{-(s+\xi)t} d\Phi(t) + \\ &\quad \varpi f(s)b(s)p_{10}^*(s) \int_0^{\infty} e^{-st} \Phi(t) dF(t) + \\ &\quad (1-\varpi) \frac{f(s)b(s)\phi(s+\xi)}{1-v(s+\xi)} \sum_{n=1}^{\infty} p_{n0}^*(s) \int_0^{\infty} e^{-(s+\xi)t} \frac{(\xi t)^n}{n!} dV(t), \end{aligned} \quad (A7)$$

$$\begin{aligned} p_{i0}^*(s) &= \frac{b^i(s)[1-f(s)]}{s} + b^i(s) \sum_{n=1}^{\infty} p_{n0}^*(s) \int_0^{\infty} \frac{(\xi t)^n}{n!} e^{-(s+\xi)t} d\Phi(t) + \\ &\quad \varpi b^i(s)p_{10}^*(s) \int_0^{\infty} e^{-st} \Phi(t) dF(t) + \\ &\quad (1-\varpi) \frac{b^i(s)\phi(s+\xi)}{1-v(s+\xi)} \sum_{n=1}^{\infty} p_{n0}^*(s) \int_0^{\infty} e^{-(s+\xi)t} \frac{(\xi t)^n}{n!} dV(t). \end{aligned} \quad (A8)$$

By solving Equations (A7) and (A8), the following relationship can be obtained

$$p_{i0}^*(s) = \frac{b^i(s)}{f(s)b(s)} \left\{ p_{00}^*(s) - \frac{1-f(s)}{s} \right\}, \quad i = 1, 2, \dots \quad (\text{A9})$$

Substituting Equation (A9) into Equation (A7) and simplifying yields Equation (11). Then, substituting Equation (11) into Equation (A9) gives Equation (12). Thus, the proof of Theorem 1 is concluded.

B Proof of Theorem 2

For $j = 1, 2, \dots$, the necessary and sufficient condition for the queue length to be j at time t is that the system is either in a busy period with queue length j , or in a non-busy period with queue length j . By applying the total probability decomposition technique, we obtain

$$\begin{aligned} p_{0j}(t) &= P\{\hat{\tau}_1 \leq t < \hat{\tau}_1 + b_1, N(t) = j\} + P\{t \geq \hat{\tau}_1 + b_1, N(t) = j\} \\ &= \int_0^t Q_j(t-x) dF(x) + P\{t \geq \hat{\tau}_1 + b_1, N(t) = j\}. \end{aligned} \quad (\text{A1})$$

The second term of Equation (A1) represents the system dynamics during the server's idle phase following the completion of a busy period. Depending on whether customers arrive during the preventive maintenance period and on the server's decision when the system becomes empty, this term can be classified into five distinct cases. These cases describe all possible system states associated with preventive maintenance, standby waiting, and multiple vacation operations.

Case 1: The system is undergoing preventive maintenance at time t , and the current queue length is observed to be j .

$$\begin{aligned} &P\{\hat{\tau}_1 + b_1 + \hat{\tau}_1 \leq t < \hat{\tau}_1 + b_1 + \Phi, 0 \leq \hat{\tau}_2 \leq \Phi, \\ &\quad \hat{\tau}_1 + b_1 + \hat{\tau}_2 + l_{j-1} \leq t < \hat{\tau}_1 + b_1 + \hat{\tau}_2 + l_j\} \\ &= \int_0^t \bar{\Phi}(t-x) \frac{[\xi(t-x)]^j}{j!} e^{-\xi(t-x)} d[F(x) * B(x)]. \end{aligned}$$

Case 2: Customers arrive during the preventive maintenance period Φ , and the time t falls within the subsequent busy period. The queue length at this moment is j . This situation is shown in Figure A4.

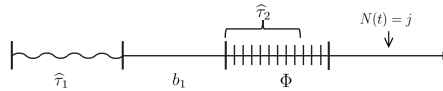


Figure A4 Development process of system under Case 2

We have

$$\begin{aligned} &\sum_{n=1}^{\infty} P\{\hat{\tau}_1 + b_1 + \hat{\tau}_2 \leq t, 0 \leq \hat{\tau}_2 \leq \Phi, \hat{\tau}_2 + l_{n-1} \leq \Phi < \hat{\tau}_2 + l_n, N(t) = j\} \\ &= \sum_{n=1}^{\infty} \int_0^t P\{\hat{\tau}_2 \leq t-x, 0 \leq \hat{\tau}_2 \leq \Phi, \hat{\tau}_2 + l_{n-1} \leq \Phi < \hat{\tau}_2 + l_n, N(t-x) = j\} d[F(x) * B(x)] \end{aligned}$$

$$\begin{aligned}
 &= \sum_{n=1}^{\infty} \int_0^t \int_0^{t-x} P\{\hat{\tau}_2 \leq t-x, \hat{\tau}_2 + l_{n-1} \leq y < \hat{\tau}_2 + l_n, N(t-x-y) = j\} d\Phi(y) d[F(x) * B(x)] \\
 &= \sum_{n=1}^{\infty} \int_0^t \int_0^{t-x} p_{nj}(t-x-y) \frac{(\xi y)^n}{n!} e^{-\xi y} d\Phi(y) d[F(x) * B(x)].
 \end{aligned}$$

Case 3: No customers arrive during the preventive maintenance period Φ , and the server chooses, with probability ϖ , to remain in the system and wait for the first arriving customer. Time t falls within the subsequent busy period, and the queue length at this moment is j . This scenario is shown in Figure A5.

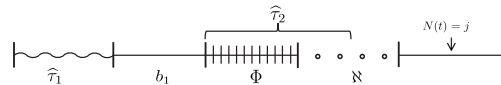


Figure A5 Development process of system under Case 3

We have

$$\begin{aligned}
 &\varpi \cdot P\{\hat{\tau}_1 + b_1 + \hat{\tau}_2 \leq t, \hat{\tau}_2 > \Phi, N(t) = j\} \\
 &= \varpi \int_0^t \int_0^{t-x} p_{1j}(t-x-y) \Phi(y) dF(y) d[F(x) * B(x)].
 \end{aligned}$$

Case 4: No customers arrive during the preventive maintenance period Φ . The server then takes multiple vacations with probability $1 - \varpi$, and the time t falls within the k th vacation V_k . The queue length at this moment is j . This scenario is illustrated in Figure A6.

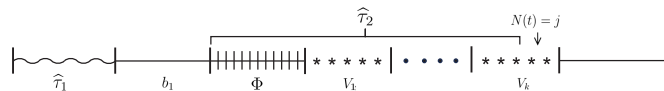


Figure A6 Development process of system under Case 4

We have

$$\begin{aligned}
 &(1 - \varpi) \sum_{k=1}^{\infty} P\{\hat{\tau}_1 + b_1 + \hat{\tau}_2 \leq t < \hat{\tau}_1 + b_1 + \Phi + s_k, \Phi + s_{k-1} \leq \hat{\tau}_2 < \Phi + s_k, \\
 &\quad \hat{\tau}_2 + l_{j-1} \leq \Phi + s_k < \hat{\tau}_2 + l_j, N(t) = j\} \\
 &= (1 - \varpi) \sum_{k=1}^{\infty} \int_0^t P\{\hat{\tau}_2 \leq t-x < \Phi + s_k, \Phi + s_{k-1} \leq \hat{\tau}_2 < \Phi + s_k, \\
 &\quad \hat{\tau}_2 + l_{j-1} \leq \Phi + s_k < \hat{\tau}_2 + l_j, N(t-x) = j\} d[F(x) * B(x)] \\
 &= (1 - \varpi) \sum_{k=1}^{\infty} \int_0^t \int_0^{t-x} P\{\hat{\tau}_2 \leq t-x-y < V_k, \hat{\tau}_2 < V_k, \hat{\tau}_2 + l_{j-1} \leq V_k < \hat{\tau}_2 + l_j, \\
 &\quad N(t-x-y) = j\} e^{-\xi y} d[\Phi(y) * V^{(k-1)}(y)] d[F(x) * B(x)] \\
 &= (1 - \varpi) \sum_{k=1}^{\infty} \int_0^t \int_0^{t-x} \bar{V}(t-x-y) \frac{[\xi(t-x-y)]^j}{j!} e^{-\xi y}
 \end{aligned}$$

$$d[\Phi(y) * V^{(k-1)}(y)]d[F(x) * B(x)],$$

where $\hat{\tau}_2$ denotes the residual time of $\hat{\tau}_2$.

Case 5: No customers arrive during the preventive maintenance period Φ . The server then initiates multiple vacations with probability $1 - \varpi$, and the time t falls into the subsequent busy period. This scenario is illustrated in Figure A7.

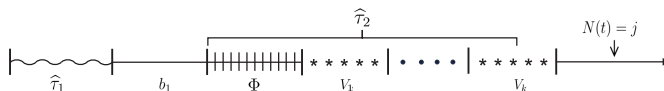


Figure A7 Development process of system under Case 5

We have

$$\begin{aligned} & (1 - \varpi) \sum_{k=1}^{\infty} \sum_{n=1}^{\infty} P\{\hat{\tau}_1 + b_1 + \Phi + s_k \leq t, \Phi + s_{k-1} \leq \hat{\tau}_2 < \Phi + s_k, \\ & \hat{\tau}_2 + l_{n-1} \leq \Phi + s_k < \hat{\tau}_2 + l_n, N(t) = j\} \\ & = (1 - \varpi) \sum_{k=1}^{\infty} \sum_{n=1}^{\infty} \int_0^t P\{\Phi + s_k \leq t - x, \Phi + s_{k-1} \leq \hat{\tau}_2 < \Phi + s_k, \\ & \hat{\tau}_2 + l_{n-1} \leq \Phi + s_k < \hat{\tau}_2 + l_n, N(t - x) = j\} d[F(x) * B(x)] \\ & = (1 - \varpi) \sum_{k=1}^{\infty} \sum_{n=1}^{\infty} \int_0^t \int_0^{t-x} P\{V_k \leq t - x - y, \hat{\tau}_2 < V_k, \\ & \hat{\tau}_2 + l_{n-1} \leq V_k < \hat{\tau}_2 + l_n, N(t - x - y) = j\} d[\Phi(y) * V^{(k-1)}(y)] d[F(x) * B(x)] \\ & = (1 - \varpi) \sum_{k=1}^{\infty} \sum_{n=1}^{\infty} \int_0^t \int_0^{t-x} \int_0^{t-x-y} p_{nj}(t - x - y - u) \frac{(\xi u)^j}{j!} e^{-\xi(u+y)} \\ & dV(u) d[\Phi(y) * V^{(k-1)}(y)] d[F(x) * B(x)]. \end{aligned}$$

By collecting the expressions above, we obtain

$$\begin{aligned} p_{0j}(t) &= \int_0^t Q_j(t - x) dF(x) + \\ & \sum_{n=1}^{\infty} \int_0^t \int_0^{t-x} p_{nj}(t - x - y) \frac{(\xi y)^n}{n!} e^{-\xi y} d\Phi(y) d[F(x) * B(x)] + \\ & \varpi \int_0^t \int_0^{t-x} p_{1j}(t - x - y) \Phi(y) dF(y) d[F(x) * B(x)] + \\ & (1 - \varpi) \sum_{k=1}^{\infty} \int_0^t \int_0^{t-x} \bar{V}(t - x - y) \frac{[\xi(t - x - y)]^j}{j!} e^{-\xi y} \\ & d[\Phi(y) * V^{(k-1)}(y)] d[F(x) * B(x)] + \\ & (1 - \varpi) \sum_{k=1}^{\infty} \sum_{n=1}^{\infty} \int_0^t \int_0^{t-x} \int_0^{t-x-y} p_{nj}(t - x - y - u) \frac{(\xi u)^j}{j!} e^{-\xi(u+y)} \\ & dV(u) d[\Phi(y) * V^{(k-1)}(y)] d[F(x) * B(x)]. \end{aligned} \tag{A2}$$

Similarly, we obtain

$$\begin{aligned}
 p_{ij}(t) = & \int_0^t Q_j(t-x) dB^{(i-1)}(x) + \\
 & \sum_{n=1}^{\infty} \int_0^t \int_0^{t-x} p_{nj}(t-x-y) \frac{(\xi y)^n}{n!} e^{-\xi y} d\Phi(y) dB^{(i)}(x) + \\
 & \varpi \int_0^t \int_0^{t-x} p_{1j}(t-x-y) \Phi(y) dF(y) dB^{(i)}(x) + \\
 & (1-\varpi) \sum_{k=1}^{\infty} \int_0^t \int_0^{t-x} \bar{V}(t-x-y) \frac{[\xi(t-x-y)]^j}{j!} e^{-\xi y} \\
 & d[\Phi(y) * V^{(k-1)}(y)] dB^{(i)}(x) + \\
 & (1-\varpi) \sum_{k=1}^{\infty} \sum_{n=1}^{\infty} \int_0^t \int_0^{t-x} \int_0^{t-x-y} p_{nj}(t-x-y-u) \frac{(\xi u)^j}{j!} e^{-\xi(u+y)} \\
 & dV(u) d[\Phi(y) * V^{(k-1)}(y)] dB^{(i)}(x). \tag{A3}
 \end{aligned}$$

Each of these cases contributes to the total response time distribution through different conditional probabilities. Equations (A2) and (A3) describe the system evolution under these conditions. Taking the LT of equations (A2) and (A3), respectively, we obtain

$$\begin{aligned}
 p_{0j}^*(s) = & f(s)q_j^*(s) + f(s)b(s) \int_0^{\infty} e^{-(s+\xi)t} \bar{\Phi}(t) \frac{(\xi t)^j}{j!} dt + \\
 & \sum_{n=1}^{\infty} f(s)b(s)p_{nj}^*(s) \int_0^{\infty} e^{-(s+\xi)t} \frac{(\xi t)^n}{n!} d\Phi(t) + \\
 & \varpi f(s)b(s)p_{1j}^*(s) \int_0^{\infty} e^{-st} \Phi(t) dF(t) + \\
 & (1-\varpi) \frac{f(s)b(s)\phi(s+\xi)}{1-v(s+\xi)} \int_0^{\infty} e^{-(s+\xi)t} \bar{V}(t) \frac{(\xi t)^j}{j!} dt + \\
 & (1-\varpi) \frac{f(s)b(s)\phi(s+\xi)}{1-v(s+\xi)} \sum_{n=1}^{\infty} p_{nj}^*(s) \int_0^{\infty} e^{-(s+\xi)t} \frac{(\xi t)^n}{n!} dV(t), \tag{A4}
 \end{aligned}$$

$$\begin{aligned}
 p_{ij}^*(s) = & \sum_{k=1}^i q_{j-i+k}^*(s) b^{k-1}(s) + b^i(s) \int_0^{\infty} e^{-(s+\xi)t} \bar{\Phi}(t) \frac{(\xi t)^j}{j!} dt + \\
 & \sum_{n=1}^{\infty} b^i(s) p_{nj}^*(s) \int_0^{\infty} e^{-(s+\xi)t} \frac{(\xi t)^n}{n!} d\Phi(t) + \\
 & \varpi b^i(s) p_{1j}^*(s) \int_0^{\infty} e^{-st} \Phi(t) dF(t) + \\
 & (1-\varpi) \frac{b^i(s)\phi(s+\xi)}{1-v(s+\xi)} \int_0^{\infty} e^{-(s+\xi)t} \bar{V}(t) \frac{(\xi t)^j}{j!} dt + \\
 & (1-\varpi) \frac{b^i(s)\phi(s+\xi)}{1-v(s+\xi)} \sum_{n=1}^{\infty} p_{nj}^*(s) \int_0^{\infty} e^{-(s+\xi)t} \frac{(\xi t)^n}{n!} dV(t). \tag{A5}
 \end{aligned}$$

From Equations (A4) and (A5), the following relationship between $p_{0j}^*(s)$ and $p_{ij}^*(s)$ is obtained

$$p_{ij}^*(s) = \sum_{k=1}^i q_{j-i+k}^*(s) b^{k-1}(s) + \frac{b^{i-1}(s)}{f(s)} \{p_{0j}^*(s) - q_j^*(s) f(s)\}, \quad i, j = 1, 2, \dots \quad (\text{A6})$$

Substituting Equation (A6) into Equation (A4) yields Equation (13). Moreover, substituting (13) into Equation (A5) yields Equation (14). Thus, this completes the proof of Theorem 2.