mathematics

MDPI

Article

Analysis of a Two-Stage Tandem Queuing System with Priority and Clearing Service in the Second Stage

Jia Xu and Liwei Liu *

School of Mathematics and Statistics, Nanjing University of Science and Technology, Nanjing 210094, China; xjia@njust.edu.cn
* Correspondence: lwliu@njust.edu.cn

Abstract: This paper considers a two-stage tandem queuing system with ordinary customers and priority customers. Upon arrival, ordinary customers are individually served in the first stage, then move to the second stage and receive clearing service. Priority customers can bypass the first stage and proceed directly to the second stage for clearing service. The second stage has N service seats. All customers currently in the second stage are served simultaneously (i.e., clearing service). Once there are N customers in the second stage, the first stage will be blocked, and newly arriving priority customers will balk and leave without joining. We first formulate a two-dimensional Markov chain to analyze this queuing system and derive the stability condition. Subsequently, the stationary distribution of the system is derived using the matrix-analytic method and spectral expansion technique. Furthermore, analytical expressions for the mean queue length, mean sojourn time, and other performance measures are presented. Finally, some numerical examples are provided to illustrate the effects of various parameters, offering valuable insights for designing such two-stage tandem queuing systems.

Keywords: two-stage service system; tandem queue; priority customers; clearing service; sojourn time

MSC: 60J28; 60K30; 90B22

Citation: Xu, J.; Liu, L. Analysis of a Two-Stage Tandem Queuing System with Priority and Clearing Service in the Second Stage. *Mathematics* **2024**, *12*, 1500. https://doi.org/10.3390/math12101500

Academic Editors: Oleg Tikhonenko and Marcin Ziółkowski

Received: 28 April 2024
Revised: 8 May 2024
Accepted: 9 May 2024
Published: 11 May 2024

1. Introduction

This study investigates a two-stage tandem queuing system with two types of customers: ordinary and priority customers. Upon arrival, an ordinary customer enters the first stage and receives a single service. After completing this service, he/she moves to the second stage and receives a clearing service. However, an arriving priority customer can bypass the first queue and directly join the second stage to receive the clearing service. The second stage has a maximum capacity of N service seats. Priority customers can only enter the system if there are fewer than N customers already in the second stage upon their arrival; otherwise, they will leave without joining the system. Ordinary customers always enter the first stage and wait for service. Once the number of customers being served simultaneously in the second stage reaches the threshold N, the first stage stops providing service, and customers in the first stage are blocked. Until a service seat becomes available in the second stage, the service in the first stage resumes.

The analyzed two-stage tandem queuing system with priority is inspired by public transportation systems. When taking public transport (e.g., buses, subways, trains, or planes), passengers holding prepaid cards, ride cards, or those who buy tickets online are given priority. They can skip the physical queue at the ticket office and directly access the waiting room. As shown in Figure 1, ordinary passengers initially proceed to the ticket office, join a queue to purchase tickets, and then move to the waiting room to prepare for boarding. Priority passengers who have ride cards or buy tickets online, on the other hand, enter the waiting room immediately. Once the scheduled departure time arrives, all

passengers in the waiting room, including both ordinary and priority passengers, board the designated vehicle and leave the station together. The seating capacity in cars, trains, or flights is predetermined and finite. As soon as all seats are filled, the ticket office stops selling tickets, and online ticket purchases are also suspended. This means that when the number of customers in the second stage reaches the threshold, ordinary customers in the first stage are all blocked, and priority customers balk.

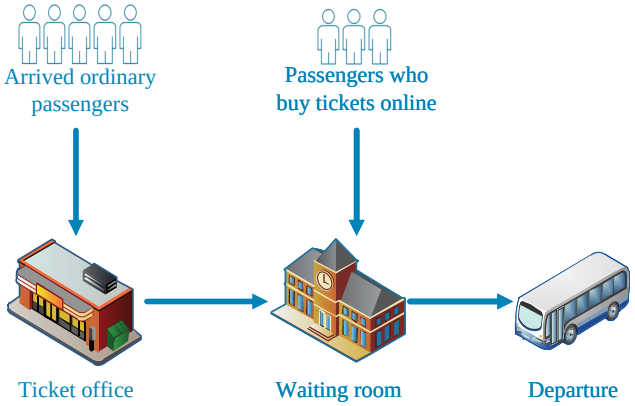


Figure 1. An example of the studied two-stage tandem queue.

Examples of the two-stage tandem service system with priority can also be seen in various sectors, such as amusement parks and production systems. In amusement parks, visitors go through a ticket check before entering the facilities. These facilities have limited capacity, and the ticket check process stops when all seats are occupied. After completing the ticket check, a group of visitors enters the facility together and experiences the services simultaneously. However, certain rides may be popular, making it challenging to enjoy them in a short amount of time. To address this, amusement parks often provide paid fast-pass tickets (e.g., Universal Studios' Express Pass and Disney's Lightning Lane). Holders of fast-pass tickets can skip the regular ticket check queue and enter the facility directly. However, if the seats are already full, those with fast-pass tickets cannot enter the facility and must wait in the fast-pass queue.

Another example is a production system: The first stage involves the factory receiving and organizing orders, while the second stage entails manufacturing products for those orders. Each batch of products can fulfill a set number of orders. Rush orders receive higher priority and are processed immediately without queuing in the first stage. The output of a factory is capped within a production cycle. When the quantity reaches the maximum, the factory ceases to accept rush orders and stops the arrangement of incoming orders.

A tandem queue is a queuing system organized in the series queue in which service facilities provide services in sequence [1–4]. Since Jackson [5] introduced the analytical frameworks of interconnected queues, tandem queuing systems have been extensively analyzed due to their applicability across various sectors (such as communication, transportation, manufacturing, networks) [6–8]. Neuts [9] studied a two-stage tandem system using an embedded semi-Markov process, where the arrival process of customers follows a Poisson process, the service time at the first stage follows a general distribution, and the service time at the second stage follows an exponential distribution, with a finite buffer between the two stages. Moreover, Neuts [10] proposed the matrix-analytic technique to study queuing systems, which helps us to handle various complex tandem queues effectively. In this paper, we employ this method to compute the steady-state distribution of the considered system.

Yang et al. [11] analyzed a two-stage tandem queuing system featuring a single service in the first stage and a batch service in the second stage, both provided by a single switched server under N policy. Recently, Nazarov et al. [12] studied a queuing network with two servers in series and two infinite orbits using the asymptotic analysis method. Dudin, Dudina, and Dudin [13] considered a dual queuing model with multiple servers in the

second stage. In their model, customers are served individually in the first stage, while services are rendered to groups with variable service time based on group size in the second stage. To the best of our knowledge, there has been no research study on such a two-stage tandem queuing system with a single service at stage one and a clearing service at stage two, where priority customers can skip stage one and join stage two immediately. The clearing service can be regarded as a flexible batch service, as it also processes multiple customers simultaneously but without limiting the batch size. This service mode tends to clear all customers in the queue at once, which is different from the fixed batch size characteristic of batch service.

In a priority queue system, customers are served based on the order of their priority. There are two types of priority in queuing systems: preemptive [14,15] and non-preemptive [16–18]. Preemptive priority occurs when a high-priority customer arrives, and the server halts service to a lower-priority customer to serve the higher-priority one immediately. Conversely, for non-preemptive priority, although newly arrived high-priority customers cannot interrupt ongoing service to lower-priority customers, they are placed ahead in the queue, or only high-priority customers are allowed to enter the system in some cases. Liu and Zhao [19] studied an $M_1, M_2/G_1, G_2/1$ queuing model with non-preemptive priority, analyzing the tail asymptotic properties of steady-state queue length. Lee et al. [20] considered the single server finite queue with non-preemptive priority and phase-type distributed service time. A two-stage tandem queue consisting of infinite servers in the first stage and finite servers in the second stage was proposed by Kim and Dudin [21]. When the number of servers occupied in the second stage reaches the threshold, non-priority customers cannot enter the system upon arrival. On the other side, Atencia [22] analyzed a discrete-time queuing model with general distributed service time and preemptive priority service, while Xie et al. [23] investigated an $M/M/1$ queuing system where preemptive service allows low-priority customers to be upgraded to high-priority. More recently, Xu, Liu, and Wu [24] considered an $M/G/1$ retrial queue with preemptive priority operating on a Bernoulli schedule, where a new arrival either immediately displaces the current customer being served with a probability α or joins the retrial orbit with a probability $1-\alpha$. Chamberlain and Starobinski [25] studied a single server queue with preemptive-resume service and general distributed service time, customer equilibrium joining strategies, and social welfare under the unobservable scenario are presented. In this paper, we study a two-stage tandem service system with priority. In contrast to the previous traditional pure preemptive service or non-preemptive service, we consider the customer who has priority to bypass the single service at stage one and immediately join stage two for clearing service.

In this study, we model the interested service system as a two-stage tandem queuing model with two types of customers, where stage one provides a single service while stage two offers a clearing service. Our focus lies on the steady-state performance analysis of such service systems. To achieve this, we investigated the system's stability condition and steady-state distribution, which are crucial for understanding its performance. Then, some system performance metrics for given parameters can be calculated using the provided computation algorithm. The main contributions and innovations of this paper are summarized as follows:

1. We propose a novel tandem queuing model consisting of a single service at stage one and a clearing service at stage two that is commonly found in real life. In the model, we also consider customers with priority to bypass stage one and directly access stage two for service, which is more in line with the practical situation.
2. We formulate the studied system as a two-dimensional Markov chain and derive the stationary distribution using the matrix-analytic and spectral expansion methods. Our theoretical results may be useful for solving problems in similar service systems.
3. We conduct sensitivity analysis of various parameters on system performance characteristics, providing a theoretical foundation and practical guidance for service system design and optimization.

The rest of this paper is organized as follows. First, we describe the queuing model and how it was formulated in Section 2. We then analyze the system stability condition in Section 3. In Section 4, we provide the system stationary probability distribution and expected sojourn time of customers. Furthermore, some essential performance measures are obtained in Section 5. In Section 6, we present the sensitivity analysis of various system parameters. Finally, the conclusion is given in Section 7.

2. Queuing Model and Its Formulation

We are interested in a two-stage tandem queuing system that includes ordinary and priority customers, where stage one provides a single service while stage two provides a clearing service. The schematic diagram of the queuing service system is illustrated in Figure 2.

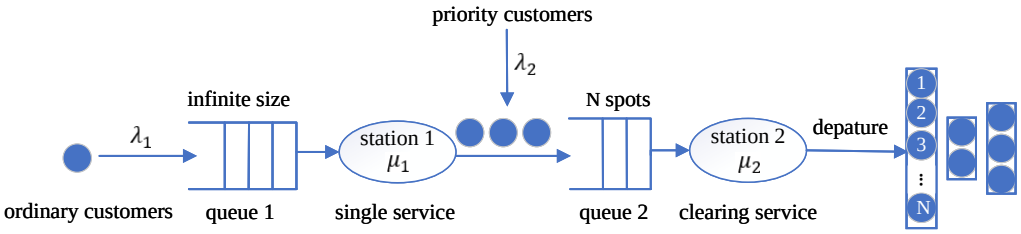


Figure 2. Two-stage tandem queuing system with single and clearing service.

The system operates with two separate queues, and each queue is managed independently by dedicated servers. Queue 1 and queue 2 represent the queues at stage one and stage two (including customers being serviced), respectively. The studied service system caters to two types of customers: (1) Ordinary customers, who first enter queue 1 and receive a single service at stage one, then move to queue 2 for clearing service at stage two; (2) Priority customers, who bypass the first stage and directly enter queue 2 for service at stage two. The detailed queuing model is explained below:

- The capacity of queue 1 is infinite, while queue 2 has N spots.
- Ordinary and priority customers arrive at the system according to two independent Poisson processes with arrival rates λ_1 and λ_2, respectively.
- Upon the arrival of a priority customer, if the number of customers in the second stage reaches the threshold N, the customer will balk (leaves without joining the system), i.e., a priority customer enters the system only when the number of customers at stage two is fewer than N upon his/her arrival.
- Ordinary customers are served individually (single service) in the first stage based on the First-Come-First-Served *(FCFS)* discipline. All customers are served together (clearing service) in the second stage. There are N customers who are being served in the second stage at most. The single service time and clearing service time are independent and exponentially distributed with parameters μ_1 and μ_2, respectively.
- If the number of customers in the second stage is fewer than N, any ordinary customer who finished their service at stage one will immediately move to stage two and receive a clearing service. However, if the queue length of the second stage reaches the capacity threshold N, the first stage and ordinary customers in the first stage will be blocked. In this case, the first stage stops providing service even if there are some customers in the first stage. The service in the first stage will be resumed only when some spots are available in the second stage.

Let $q_1(t)$ and $q_2(t)$ be the number of customers in queue 1 and queue 2 at time t (including the customers being serviced), respectively. Then, the queuing model is formulated as a two-dimensional continuous-time Markov chain $\{(q_1(t), q_2(t)), t \geq 0\}$ with state space

$$\Omega = \{(0,0),(0,1),\cdots,(0,N);(1,0),\cdots,(1,N);\cdots;(n,0),(n,1),\cdots,(n,N);\cdots\}.$$

When $q_2 = N$, the first stage stops serving customers, and the first stage is blocked, as are the customers at stage one. Please note that this case does not contribute to the probability that the first stage is busy (working). If all system parameters are positive, it is clear that the continuous-time Markov chain $\{(q_1(t), q_2(t)), t \geq 0\}$ is irreducible.

Let $q((q_1, q_2), (y_1, y_2))$ be the transition rate from state (q_1, q_2) to state (y_1, y_2), then the infinitesimal generator of the Markov chain is $\boldsymbol{Q} = (q((q_1, q_2), (y_1, y_2)))$, $(q_1, q_2) \in \Omega$, $(y_1, y_2) \in \Omega$. The transition rate diagram for this two-stage tandem service system is depicted in Figure 3. For $(q_1, q_2) \in \Omega$, the one-stop transition rates are given by

$$\begin{aligned} q\big((q_1, q_2), (q_1 + 1, q_2)\big) &= \lambda_1, \\ q\big((q_1, q_2), (q_1, q_2 + 1)\big) &= \lambda_2, \\ q\big((q_1, q_2), (q_1 - 1, q_2 + 1)\big) &= \mu_1, \ (q_1 > 0, \ q_2 < N), \\ q\big((q_1, q_2), (q_1, 0)\big) &= \mu_2, \ (q_2 > 0). \end{aligned}$$

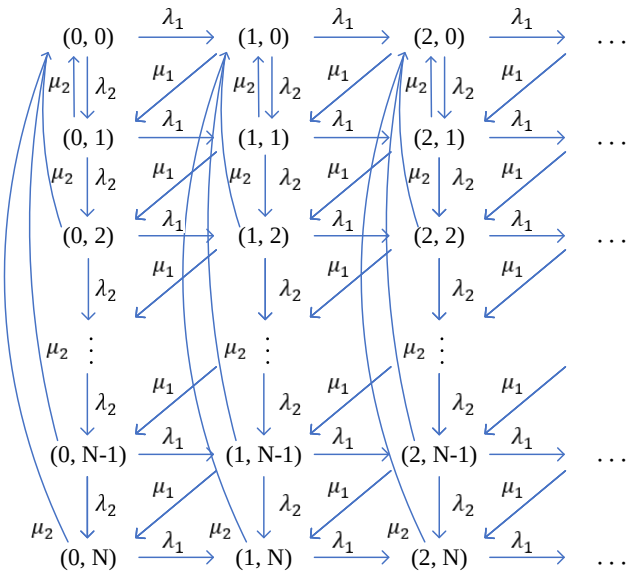


Figure 3. State-transition diagram of the queuing system.

3. Stability Condition

This section analyzes the system stability condition based on the matrix-geometric solution in [26]. The infinitesimal generator of the above continuous-time Markov chain $\{(q_1(t), q_2(t)), t \geq 0\}$ can be written as the block-partitioned form:

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{B} & \boldsymbol{A_0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \boldsymbol{A_2} & \boldsymbol{A_1} & \boldsymbol{A_0} & \boldsymbol{0} & \boldsymbol{0} & \cdots \\ \boldsymbol{0} & \boldsymbol{A_2} & \boldsymbol{A_1} & \boldsymbol{A_0} & \boldsymbol{0} & \cdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \end{pmatrix},$$

the matrices $\boldsymbol{B}$, $\boldsymbol{A_0}$, $\boldsymbol{A_1}$, and $\boldsymbol{A_2}$ are both $(N+1)$-dimensional square matrices with elements, where

$$\boldsymbol{B} = \begin{pmatrix} -(\lambda_1 + \lambda_2) & \lambda_2 & & & \\ \mu_2 & -(\lambda_1 + \lambda_2 + \mu_2) & \lambda_2 & & \\ \vdots & & \ddots & \ddots & \\ \mu_2 & & & -(\lambda_1 + \lambda_2 + \mu_2) & \lambda_2 \\ \mu_2 & & & & -(\lambda_1 + \mu_2) \end{pmatrix},$$

$$\boldsymbol{A_0} = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_1 & & \\ & & \ddots & \\ & & & \lambda_1 \end{pmatrix}, \ \boldsymbol{A_2} = \begin{pmatrix} 0 & \mu_1 & & \\ & 0 & \ddots & \\ & & \ddots & \mu_1 \\ & & & 0 \end{pmatrix},$$

$$\boldsymbol{A_1} = \begin{pmatrix} -(\lambda_1+\lambda_2+\mu_1) & \lambda_2 & & & & \\ \mu_2 & -(\lambda_1+\lambda_2+\mu_1+\mu_2) & \lambda_2 & & & \\ \vdots & & \ddots & \ddots & & \\ \mu_2 & & & -(\lambda_1+\lambda_2+\mu_1+\mu_2) & \lambda_2 & \\ \mu_2 & & & & -(\lambda_1+\mu_2) \end{pmatrix}.$$

Obviously, according to the characteristics of matrix $\boldsymbol{Q}$, we can know that the Markov chain $\{(q_1(t), q_2(t)), t \geq 0\}$ is a quasi-birth-and-death (QBD) process.

Theorem 1. *The considered two-stage tandem queuing system is stable if and only if*

$$\rho = \frac{\lambda_1}{\mu_1} + \left(\frac{\lambda_2+\mu_1}{\lambda_2+\mu_1+\mu_2}\right)^N < 1.$$

Proof. According to the mean drift result in [10], the system would be stable, and the stationary distributions exist if and only if

$$\boldsymbol{X}\boldsymbol{A_0}\boldsymbol{e} < \boldsymbol{X}\boldsymbol{A_2}\boldsymbol{e},$$

where $\boldsymbol{e}$ is a $(N+1)$-dimensional column vector with all elements are equal to 1, $\boldsymbol{X} = (X_1, X_2, \cdots, X_{N+1})$ is the invariant probability vector of $\boldsymbol{A} = \boldsymbol{A_0} + \boldsymbol{A_1} + \boldsymbol{A_2}$, and

$$\boldsymbol{A} = \begin{matrix} 1 \\ 2 \\ \vdots \\ N \\ N+1 \end{matrix} \begin{pmatrix} -(\lambda_2+\mu_1) & \lambda_2+\mu_1 & & & \\ \mu_2 & -(\lambda_2+\mu_1+\mu_2) & \lambda_2+\mu_1 & & \\ \vdots & & \ddots & \ddots & \\ \mu_2 & & & -(\lambda_2+\mu_1+\mu_2) & \lambda_2+\mu_1 \\ \mu_2 & & & & -\mu_2 \end{pmatrix}.$$

Then, we have the following balance equations

$$\boldsymbol{X}\boldsymbol{A} = \boldsymbol{0}, \tag{1}$$

$$\boldsymbol{X}\boldsymbol{e} = 1. \tag{2}$$

Expanding Equation (1) yields the following equations

$$\begin{gathered} -(\lambda_2+\mu_1)X_1 + \mu_2 \sum_{n=2}^{N+1} X_n = 0, \\ (\lambda_2+\mu_1)X_{n-1} - (\lambda_2+\mu_1+\mu_2)X_n = 0, \ 2 \leq n \leq N, \\ (\lambda_2+\mu_1)X_N - \mu_2 X_{N+1} = 0, \end{gathered}$$

which, by recursive iteration, yield

$$X_n = \begin{cases} \left(\dfrac{\lambda_2+\mu_1}{\lambda_2+\mu_1+\mu_2}\right)^{n-1} X_1, & 1 \leq n \leq N, \\ \dfrac{\lambda_2+\mu_1}{\mu_2}\left(\dfrac{\lambda_2+\mu_1}{\lambda_2+\mu_1+\mu_2}\right)^{N-1} X_1, & n = N+1. \end{cases} \tag{3}$$

Substituting (3) into (2) results in

$$X_{N+1} = \left(\frac{\lambda_2 + \mu_1}{\lambda_2 + \mu_1 + \mu_2}\right)^N. \tag{4}$$

The stability condition $\boldsymbol{XA_0e} < \boldsymbol{XA_2e}$ can be rewritten as

$$\lambda_1 < \mu_1(1 - X_{N+1}). \tag{5}$$

By substituting (4) into (5), we obtain the stability condition of the studied system

$$\lambda_1 < \mu_1\left[1 - \left(\frac{\lambda_2 + \mu_1}{\lambda_2 + \mu_1 + \mu_2}\right)^N\right]. \tag{6}$$

After transforming the above equation, we have

$$\rho = \frac{\lambda_1}{\mu_1} + \left(\frac{\lambda_2 + \mu_1}{\lambda_2 + \mu_1 + \mu_2}\right)^N < 1.$$

Thus, the proof is completed. □

Remark 1. *Due to the finite capacity of the second stage, achieving a steady-state system necessitates that the arrival rate of ordinary customers at the first stage remains below the actual service rate of the first stage. If the two queues are independent, the service rate of the first stage is μ_1. However, as the two queues are tandem in series, the departure from the first stage is constrained by the capacity of the second stage. Please note that the term $\left(\frac{\lambda_2+\mu_1}{\lambda_2+\mu_1+\mu_2}\right)^N$ represents the probability that the second stage queue reaches its maximum capacity N (i.e., the probability that the first stage is blocked). Therefore, $\mu_1\left[1 - \left(\frac{\lambda_2+\mu_1}{\lambda_2+\mu_1+\mu_2}\right)^N\right]$ means the effective service rate of the first stage. Thus, the stability condition ensures a balance between the arrival rate of ordinary customers and the ability of this system to accept and process them, considering the limitations imposed by the finite capacity of the second stage and priority customers, therefore ensuring the maintenance of stable system operation.*

4. Steady-State Analysis

Under the condition of system stability, this section will determine the stationary distribution and calculate the sojourn times of customers within the system.

4.1. Stationary Probability Distribution

This section will calculate the stationary probabilities for the queuing model under study using matrix-analytic and spectral expansion methods. We first define the steady-state joint probabilities of the Markov chain under consideration as:

$$\begin{aligned}
\pi_{i,j} &= \lim_{t\to\infty} P\{q_1(t) = i, q_2(t) = j)\},\ i \geq 0,\ 0 \leq j \leq N.\\
\boldsymbol{\pi_i} &= (\pi_{i,0}, \pi_{i,1}, \pi_{i,2}, \cdots, \pi_{i,N}),\ i \geq 0,\\
\boldsymbol{\pi} &= (\boldsymbol{\pi_0}, \boldsymbol{\pi_1}, \boldsymbol{\pi_2}, \cdots).
\end{aligned}$$

When the stability condition $\rho < 1$ is satisfied, the steady-state system has the following Chapman-Kolmogorov equations based on the above definitions:

$$\begin{aligned}
(\lambda_1 + \lambda_2)\pi_{0,0} &= \mu_2 \sum_{j=1}^{N} \pi_{0,j},\\
(\lambda_1 + \lambda_2 + \mu_2)\pi_{0,j} &= \lambda_2\pi_{0,j-1} + \mu_1\pi_{1,j-1},\ 1 \leq j \leq N-1,
\end{aligned}$$

$$
\begin{aligned}
(\lambda_1+\mu_2)\pi_{0,N} &= \lambda_2\pi_{0,N-1}+\mu_1\pi_{1,N-1},\\
(\lambda_1+\lambda_2+\mu_1)\pi_{i,0} &= \lambda_1\pi_{i-1,0}+\mu_2\sum_{j=1}^{N}\pi_{i,j},\ i\geq 1,\\
(\lambda_1+\lambda_2+\mu_1+\mu_2)\pi_{i,j} &= \lambda_1\pi_{i-1,j}+\lambda_2\pi_{i,j-1}+\mu_1\pi_{i+1,j-1},\ i\geq 1,\ 1\leq j\leq N-1,\\
(\lambda_1+\mu_2)\pi_{i,N} &= \lambda_1\pi_{i-1,N}+\lambda_2\pi_{i,N-1}+\mu_1\pi_{i+1,N-1},\ i\geq 1.
\end{aligned}
$$

The normalization equation is

$$\sum_{i=0}^{\infty}(\pi_{i,0}+\pi_{i,1}+\pi_{i,2}+\cdots+\pi_{i,N})=1.$$

From $\boldsymbol{\pi Q}=\mathbf{0}$, we have

$$\boldsymbol{\pi_0 B}+\boldsymbol{\pi_1 A_2}=\mathbf{0}, \tag{7}$$

$$\boldsymbol{\pi_{i-1}A_0}+\boldsymbol{\pi_i A_1}+\boldsymbol{\pi_{i+1}A_2}=\mathbf{0},\ i\geq 1. \tag{8}$$

Employing the matrix-geometric method described in Neuts [10], we establish that

$$\boldsymbol{\pi_i}=\boldsymbol{\pi_0 R}^i,\ i\geq 0, \tag{9}$$

where $\boldsymbol{R}$ is the minimal nonnegative solution to the matrix quadratic equation

$$\boldsymbol{A_0}+\boldsymbol{RA_1}+\boldsymbol{R^2A_2}=\mathbf{0}.$$

The matrices $\boldsymbol{A_0}$, $\boldsymbol{A_1}$, and $\boldsymbol{A_2}$ have complex forms, making direct analytical expressions for the rate matrix $\boldsymbol{R}$ challenging. The rate matrix $\boldsymbol{R}$ is typically calculated using the Gauss-Seidel iterative method, starting with

$$\boldsymbol{R_0}=\mathbf{0},$$

and updating with

$$\boldsymbol{R_{i+1}}=-\left(\boldsymbol{R_i}^2\boldsymbol{A_2}+\boldsymbol{A_0}\right)\boldsymbol{A_1}^{-1},\ i\geq 0.$$

The specifics of this approximation technique are outlined in Algorithm 1.

Algorithm 1: Pseudo code for deriving the rate matrix $\boldsymbol{R}$.

Input: Tolerance $\varepsilon>0$, $n=1$, matrix $\boldsymbol{A_0}$, $\boldsymbol{A_1}$, and $\boldsymbol{A_2}$.
Output: Rate matrix $\boldsymbol{R}$.
Step 1: Set $\boldsymbol{R_0}=\mathbf{0}$.
Step 2: $\boldsymbol{R_1}=-\left(\boldsymbol{R_0}^2\boldsymbol{A_2}+\boldsymbol{A_0}\right)\boldsymbol{A_1}^{-1}$.
Step 3: **While** $\|\boldsymbol{R_n}-\boldsymbol{R_{n-1}}\|>\varepsilon$ **do**
$n=n+1$;
$\boldsymbol{R_n}=-\left(\boldsymbol{R_{n-1}}^2\boldsymbol{A_2}+\boldsymbol{A_0}\right)\boldsymbol{A_1}^{-1}$.
End while
Step 4: $\boldsymbol{R}=\boldsymbol{R_n}$.

By substituting (9) into the first matrix balance Equation (7), it is obtained

$$\boldsymbol{\pi_0}(\boldsymbol{B}+\boldsymbol{RA_2})=\mathbf{0}, \tag{10}$$

and substituting (9) in the normalizing equation results in

$$\sum_{i=0}^{\infty}\boldsymbol{\pi_0 R}^i\boldsymbol{e}=\boldsymbol{\pi_0}(\boldsymbol{I}-\boldsymbol{R})^{-1}\boldsymbol{e}=1, \tag{11}$$

where $\boldsymbol{I}$ is a $(N+1) \times (N+1)$ identity matrix (i.e., all its diagonal elements are 1 and everywhere else is 0), and $\boldsymbol{e}$ is a $(N+1)$-dimensional column vector with all elements are 1. Then $\boldsymbol{\pi_0}$ can be obtained by solving Equations (10) and (11), and starting from $\boldsymbol{\pi_0}$, the system's stationary probability vector $(\boldsymbol{\pi_0}, \boldsymbol{\pi_1}, \boldsymbol{\pi_2}, \cdots)$ can be calculated using $\boldsymbol{\pi_i} = \boldsymbol{\pi_0 R}^i$.

Based on the analysis presented, we propose the following theorem.

Theorem 2. *When the stability condition $\rho < 1$ is satisfied, the steady-state probabilities of the system can be expressed as*

$$\boldsymbol{\pi_i} = \boldsymbol{\pi_0 R}^i, \ i \geq 0,$$

where $\boldsymbol{\pi_0}$ is determined by the unique solution to the equations

$$\begin{aligned} \boldsymbol{\pi_0}(\boldsymbol{B} + \boldsymbol{RA_2}) &= \boldsymbol{0}, \\ \boldsymbol{\pi_0}(\boldsymbol{I} - \boldsymbol{R})^{-1}\boldsymbol{e} &= 1. \end{aligned}$$

Let $E[L_1]$ be the expected queue length of the first stage (including the customer being serviced), and let $E[L_2]$ be the expected queue length of the second stage. Note that all customers in the first stage are ordinary customers. Then, the expected number of ordinary customers in the first stage is given by

$$E[L_1] = \sum_{i=0}^{\infty}\sum_{j=0}^{N} i\pi_{i,j} = \sum_{i=0}^{\infty} i\boldsymbol{\pi_i e} = \sum_{i=0}^{\infty} i\boldsymbol{\pi_0 R}^i\boldsymbol{e} = \boldsymbol{\pi_0}\left(\sum_{i=0}^{\infty} i\boldsymbol{R}^i\right)\boldsymbol{e} = \boldsymbol{\pi_0 R}(\boldsymbol{I} - \boldsymbol{R})^{-2}\boldsymbol{e}.$$

For $\boldsymbol{v} = (0, 1, 2, \cdots, N-1, N)^T$, the expected queue length of the second stage is given by

$$E[L_2] = \sum_{i=0}^{\infty}\sum_{j=0}^{N} j\pi_{i,j} = \sum_{i=0}^{\infty} \boldsymbol{\pi_i v} = \boldsymbol{\pi_0}\left(\sum_{i=0}^{\infty} \boldsymbol{R}^i\right)\boldsymbol{v} = \boldsymbol{\pi_0}(\boldsymbol{I} - \boldsymbol{R})^{-1}\boldsymbol{v}.$$

Thus, the expected number of customers in the system is given by

$$E[L] = E[L_1] + E[L_2] = \boldsymbol{\pi_0}\Big[\boldsymbol{R}(\boldsymbol{I} - \boldsymbol{R})^{-2}\boldsymbol{e} + (\boldsymbol{I} - \boldsymbol{R})^{-1}\boldsymbol{v}\Big].$$

As an alternative to the matrix-analytic method, we next use another technique (spectral expansion method) to derive the steady-state probability distribution of the studied two-stage tandem queuing system with priority, which is favorable in some cases [26,27].

The characteristic matrix polynomial of equation $\boldsymbol{\pi_{i-1}A_0} + \boldsymbol{\pi_i A_1} + \boldsymbol{\pi_{i+1}A_2} = 0$, $(i \geq 1)$ has the following structure:

$$\boldsymbol{Q}(x) = \boldsymbol{A_0} + \boldsymbol{A_1}x + \boldsymbol{A_2}x^2 = \begin{matrix} 1 \\ 2 \\ \vdots \\ N \\ N+1 \end{matrix} \begin{pmatrix} d_1(x) & d_4(x) & & & & \\ \mu_2 x & d_2(x) & d_4(x) & & & \\ \vdots & & \ddots & \ddots & & \\ \mu_2 x & & & d_2(x) & d_4(x) \\ \mu_2 x & & & & d_3(x) \end{pmatrix},$$

where

$$\begin{aligned} d_1(x) &= \lambda_1 - (\lambda_1 + \lambda_2 + \mu_1)x, \\ d_2(x) &= \lambda_1 - (\lambda_1 + \lambda_2 + \mu_1 + \mu_2)x, \\ d_3(x) &= \lambda_1 - (\lambda_1 + \mu_2)x, \\ d_4(x) &= \lambda_2 x + \mu_1 x^2. \end{aligned}$$

According to Proposition 2 in [26], if the system is stable, the characteristic polynomial $\boldsymbol{Q}(x)$ contains $N+1$ eigenvalues inside the unit circle. Consequently, the steady-state probabilities can be reformulated as follows:

$$\boldsymbol{\pi_i} = \sum_{k=1}^{N+1} a_k \nu_k^i \boldsymbol{\varphi_k},\ i \geq 0,$$

where a_k is the coefficient determined through balance equations, and the normalization condition, ν_k represents the eigenvalues within the unit circle, and $\boldsymbol{\varphi_k}$ are the corresponding left eigenvectors. A pair of $(\nu_k, \boldsymbol{\varphi_k})$ associated with $\boldsymbol{Q}(x)$ satisfy

$$\begin{aligned} \boldsymbol{\varphi_k} \boldsymbol{Q}(\nu_k) &= \mathbf{0}, \\ \det(\boldsymbol{Q}(\nu_k)) &= 0. \end{aligned}$$

Define $\nu_1, \nu_2, \cdots, \nu_{N+1}$ as the eigenvalues of $\boldsymbol{Q}(x)$ inside the unit circle, with the corresponding eigenvectors $\boldsymbol{\varphi_k} = (\varphi_{k,1}, \varphi_{k,2}, \cdots, \varphi_{k,N+1})$. From the equation $\boldsymbol{\varphi_k} \boldsymbol{Q}(\nu_k) = \mathbf{0}$, we have

$$\varphi_{k,1} d_1(\nu_k) + \sum_{j=2}^{N+1} \varphi_{k,j} \mu_2 \nu_k = 0,$$

$$\varphi_{k,j}(\lambda_2 \nu_k + \mu_1 \nu_k^2) + \varphi_{k,j+1} d_2(\nu_k) = 0,\ j = 1, 2, \cdots, N-1,$$

$$\varphi_{k,N}(\lambda_2 \nu_k + \mu_1 \nu_k^2) + \varphi_{k,N+1} d_3(\nu_k) = 0.$$

Letting $\varphi_{k,1} = 1$, after some calculation, it can be obtained that

$$\varphi_{k,j} = \begin{cases} \left(-\dfrac{\lambda_2 \nu_k + \mu_1 \nu_k^2}{d_2(\nu_k)} \right)^{j-1}, & j = 1, 2, \cdots, N, \\ -\dfrac{\lambda_2 \nu_k + \mu_1 \nu_k^2}{d_3(\nu_k)} \left(-\dfrac{\lambda_2 \nu_k + \mu_1 \nu_k^2}{d_2(\nu_k)} \right)^{N-1}, & j = N+1. \end{cases}$$

Therefore, the coefficients a_k are computed using the balance equations and normalization condition based on the derived eigenvalues and corresponding left eigenvectors:

$$\sum_{k=1}^{N+1} a_k \left[(\lambda_1 + \lambda_2) \varphi_{k,1} - \mu_2 \sum_{j=2}^{N+1} \varphi_{k,j} \right] = 0,$$

$$\sum_{k=1}^{N+1} a_k \Big[(\lambda_1 + \lambda_2 + \mu_2) \varphi_{k,j} - \lambda_2 \varphi_{k,j-1} - \mu_1 \nu_k \varphi_{k,j-1} \Big] = 0,\ j = 2, 3, \cdots, N,$$

$$\sum_{k=1}^{N+1} a_k [(\lambda_1 + \mu_2) \varphi_{k,N+1} - \lambda_2 \varphi_{k,N} - \mu_1 \nu_k \varphi_{k,N}] = 0,$$

$$\sum_{k=1}^{N+1} a_k \frac{1}{1 - \nu_k} \boldsymbol{\varphi_k} \boldsymbol{e} = 1.$$

Note that the system comprises $N+1$ independent linear equations, which facilitate the determination of the $N+1$ unknown coefficients $a_1, a_2, \cdots, a_N + 1$. Once these coefficients are derived, explicit expressions for the stationary probabilities can be obtained.

4.2. Sojourn Time of a Customer in the System

Define W as the expected sojourn time of an ordinary customer in the system and T as the expected sojourn time of a priority customer in the system. For every priority customer, the time spent in the second stage is equivalent to their total time in the system. Since customers in the second stage are served simultaneously, the expected residual service time in the second stage corresponds to the average time a customer spends in this stage. Additionally, the clearing service time follows an exponential distribution, which is

characterized by its memoryless property. Thus, the expected sojourn time of a priority customer in the system is

$$T = \frac{1}{\mu_2}.$$

Define $W_{i,j}(i \geq 0, 0 \leq j \leq N)$ as the conditional expected sojourn time of an ordinary customer, assuming there are i customers in the first stage and j customers in the second stage when the ordinary customer arrives. In the following, Theorem 3 provides the specific expression for computing the expected sojourn time of ordinary customers $W_{i,j}$.

Theorem 3. *If the stability condition $\rho < 1$ is satisfied, and the system is in state (i, j) when an ordinary customer arrives, the expected sojourn time of the ordinary customer in the system can be given as follows,*

$$W_{i,0} = \frac{\mu_2 f_{i,N} - 1}{\mu_2(1 - h_N)},$$
$$W_{i,j} = f_{i,j} + h_j W_{i,0},\ 1 \leq j \leq N,$$

where

$$h_j = \begin{cases} \dfrac{\mu_1 + \lambda_2}{\lambda_2}, & j = 1, \\ \dfrac{(\mu_1 + \lambda_2 + \mu_2)h_{j-1} - \mu_2}{\lambda_2}, & 2 \leq j \leq N, \end{cases}$$

$$f_{i,j} = \begin{cases} -\dfrac{\mu_1 + \mu_2}{\lambda_2 \mu_2}, & i = 0,\ j = 1, \\ f_{0,1} + \dfrac{\mu_1 + \lambda_2 + \mu_2}{\lambda_2} f_{0,j-1}, & i = 0,\ 2 \leq j \leq N, \\ -\dfrac{1 + \mu_1 W_{i-1,1}}{\lambda_2}, & i \geq 1,\ j = 1, \\ -\dfrac{1 + \mu_1 W_{i-1,j}}{\lambda_2} + \dfrac{\mu_1 + \lambda_2 + \mu_2}{\lambda_2} f_{i,j-1}, & i \geq 1,\ 2 \leq j \leq N. \end{cases}$$

Proof. Case (a): $i = 0$. There are no customers in the first stage upon the arrival of a new ordinary customer. By conditioning on the next future event (first-step analysis) and using the strong Markov property, we have

$$W_{0,0} = \frac{1}{\mu_1 + \lambda_2} + \frac{\mu_1}{\mu_1 + \lambda_2}\frac{1}{\mu_2} + \frac{\lambda_2}{\mu_1 + \lambda_2} W_{0,1}, \tag{12}$$

$$W_{0,j} = \frac{1}{\mu_1 + \lambda_2 + \mu_2} + \frac{\mu_1}{\mu_1 + \lambda_2 + \mu_2}\frac{1}{\mu_2} + \frac{\lambda_2}{\mu_1 + \lambda_2 + \mu_2} W_{0,j+1} + \frac{\mu_2}{\mu_1 + \lambda_2 + \mu_2} W_{0,0},\ 1 \leq j \leq N - 1, \tag{13}$$

$$W_{0,N} = \frac{1}{\mu_2} + W_{0,0}. \tag{14}$$

From (12) and (13), we can obtain

$$W_{0,1} = -\frac{\mu_1 + \mu_2}{\lambda_2 \mu_2} + \frac{\mu_1 + \lambda_2}{\lambda_2} W_{0,0},$$

$$W_{0,2} = -\frac{\mu_1 + \mu_2}{\lambda_2 \mu_2}\left(1 + \frac{\mu_1 + \lambda_2 + \mu_2}{\lambda_2}\right) + \left(\frac{\mu_1 + \lambda_2 + \mu_2}{\lambda_2}\frac{\mu_1 + \lambda_2}{\lambda_2} - \frac{\mu_2}{\lambda_2}\right) W_{0,0}.$$

Let

$$f_{0,1} = -\frac{\mu_1 + \mu_2}{\lambda_2 \mu_2},\ h_1 = \frac{\mu_1 + \lambda_2}{\lambda_2},$$

which yields

$$W_{0,1} = f_{0,1} + h_1 W_{0,0}.$$

Then, we have

$$W_{0,2} = f_{0,2} + h_2 W_{0,0},$$

where

$$f_{0,2} = f_{0,1} + \frac{\mu_1 + \lambda_2 + \mu_2}{\lambda_2} f_{0,1},\ h_2 = \frac{(\mu_1 + \lambda_2 + \mu_2)h_1 - \mu_2}{\lambda_2}.$$

Similarly, for $j = 2, 3, 4, \cdots, N$, the following recursion holds

$$W_{0,j} = f_{0,j} + h_j W_{0,0},$$

where

$$f_{0,j} = f_{0,1} + \frac{\mu_1 + \lambda_2 + \mu_2}{\lambda_2} f_{0,j-1},\ h_j = \frac{(\mu_1 + \lambda_2 + \mu_2)h_{j-1} - \mu_2}{\lambda_2}.$$

Substituting $W_{0,N} = f_{0,N} + h_N W_{0,0}$ into (14), which results in

$$W_{0,0} = \frac{\mu_2 f_{0,N} - 1}{\mu_2(1 - h_N)},$$

thus, all $W_{0,j} (j = 1, 2, \cdots, N)$ can be derived.

Case (b): $i = 1$. There are i customers in the first stage upon the arrival of a new ordinary customer. Then, the conditional expected sojourn time of an ordinary customer is as follows

$$W_{1,0} = \frac{1}{\mu_1 + \lambda_2} + \frac{\mu_1}{\mu_1 + \lambda_2} W_{0,1} + \frac{\lambda_2}{\mu_1 + \lambda_2} W_{1,1}, \tag{15}$$

$$W_{1,j} = \frac{1}{\mu_1 + \lambda_2 + \mu_2} + \frac{\mu_1}{\mu_1 + \lambda_2 + \mu_2} W_{0,j+1} + \frac{\lambda_2}{\mu_1 + \lambda_2 + \mu_2} W_{1,j+1} + \frac{\mu_2}{\mu_1 + \lambda_2 + \mu_2} W_{1,0},\ 1 \le j \le N - 1, \tag{16}$$

$$W_{1,N} = \frac{1}{\mu_2} + W_{1,0}. \tag{17}$$

From Equation (15), it arrives at

$$W_{1,1} = -\frac{1 + \mu_1 W_{0,1}}{\lambda_2} + \frac{\mu_1 + \lambda_2}{\lambda_2}, W_{1,0}.$$

Let

$$f_{1,1} = -\frac{1 + \mu_1 W_{0,1}}{\lambda_2},$$

then the above equation can be rewritten as

$$W_{1,1} = f_{1,1} + h_1 W_{1,0}.$$

Similarly, when $j = 2, 3, \cdots, N$, from Equation (16), we can obtain

$$W_{1,j} = f_{1,j} + h_j W_{1,0},$$

where

$$f_{1,j} = -\frac{1 + \mu_1 W_{0,j}}{\lambda_2} + \frac{\mu_1 + \lambda_2 + \mu_2}{\lambda_2} f_{1,j-1},\ 2 \le j \le N.$$

Therefore, it can be directly obtained

$$W_{1,N} = f_{1,N} + h_N W_{1,0},$$

combining with Equation (17), yields

$$W_{1,0} = \frac{\mu_2 f_{1,N} - 1}{\mu_2(1 - h_N)}.$$

Then, all conditional expected sojourn time $W_{1,j}(1 \le j \le N)$ have been obtained.

Case (c): $i > 1$. In the same line as Case (a) and Case (b), for a given i, the conditional expected sojourn time of an ordinary customer $W_{i,j}(0 \le j \le N)$ can be given by

$$\begin{aligned}
W_{i,0} &= \frac{1}{\mu_1 + \lambda_2} + \frac{\mu_1}{\mu_1 + \lambda_2} W_{i-1,1} + \frac{\lambda_2}{\mu_1 + \lambda_2} W_{i,1}, \\
W_{i,j} &= \frac{1}{\mu_1 + \lambda_2 + \mu_2} + \frac{\mu_1}{\mu_1 + \lambda_2 + \mu_2} W_{i-1,j+1} + \frac{\lambda_2}{\mu_1 + \lambda_2 + \mu_2} W_{i,j+1} \\
&\quad + \frac{\mu_2}{\mu_1 + \lambda_2 + \mu_2} W_{i,0},\ 1 \le j \le N-1, \\
W_{i,N} &= \frac{1}{\mu_2} + W_{i,0}.
\end{aligned}$$

From the above equations, we have the following results after some recursive calculations

$$\begin{aligned}
W_{i,0} &= \frac{\mu_2 f_{i,N} - 1}{\mu_2(1 - h_N)}, \\
W_{i,j} &= f_{i,j} + h_j W_{i,0},\ 1 \le j \le N,
\end{aligned}$$

where

$$\begin{aligned}
f_{i,1} &= -\frac{1 + \mu_1 W_{i-1,1}}{\lambda_2}, \\
f_{i,j} &= -\frac{1 + \mu_1 W_{i-1,j}}{\lambda_2} + \frac{\mu_1 + \lambda_2 + \mu_2}{\lambda_2} f_{i,j-1},\ 2 \le j \le N.
\end{aligned}$$

Then, all the expected sojourn time $W_{i,j}$ are obtained. □

According to the total probability theorem, the expected sojourn time of an ordinary customer in the system is given by

$$W = \sum_{i=0}^{\infty} \pi_{i,0} W_{i,0} + \sum_{i=0}^{\infty} \sum_{j=1}^{N} \pi_{i,j} W_{i,j}.$$

5. Performance Measures

Let $\boldsymbol{f} = (1, 0 \cdots, 0)^T$, $\boldsymbol{g} = (1, \cdots, 1, 0)^T$, $\boldsymbol{h} = (0, 1, \cdots, 1)^T$, and $\boldsymbol{u} = (0, \cdots, 0, 1)^T$, where the dimensions of these vectors should be clear from the context. In the following, some other system performance characteristics are presented in the form of steady-state probabilities.

The probability that the first stage is empty is

$$P_{e1} = \sum_{j=0}^{N} \pi_{0,j} = \boldsymbol{\pi_0 e}.$$

The probability that the second stage is empty is

$$P_{e2} = \sum_{i=0}^{\infty} \pi_{i,0} = \boldsymbol{\pi_0}(\boldsymbol{I} - \boldsymbol{R})^{-1}\boldsymbol{f}.$$

The probability that the system is empty is

$$P_e = \pi_{0,0} = \boldsymbol{\pi_0}\boldsymbol{f}.$$

The probability that the first stage is busy is

$$P_{b1} = \sum_{i=1}^{\infty}\sum_{j=0}^{N-1} \pi_{i,j} = \left[\boldsymbol{\pi_0}(\boldsymbol{I} - \boldsymbol{R})^{-1} - \boldsymbol{\pi_0}\right]\boldsymbol{g}.$$

The probability that the second stage is busy is

$$P_{b2} = \sum_{i=0}^{\infty}\sum_{j=1}^{N} \pi_{i,j} = \boldsymbol{\pi_0}(\boldsymbol{I} - \boldsymbol{R})^{-1}\boldsymbol{h}.$$

The probability that the second stage operates with the full load is

$$P_f = \sum_{i=0}^{\infty} \pi_{i,N} = \boldsymbol{\pi_0}(\boldsymbol{I} - \boldsymbol{R})^{-1}\boldsymbol{u}.$$

6. Numerical Examples

In this section, some numerical examples are presented based on previous theoretical results. All parameters used for the figures and tables meet the system stability condition. We investigate the effects of system parameters $\{\lambda_1, \lambda_2, \mu_1, \mu_2, N\}$ on mean queue length, expected sojourn time of ordinary customers, and other essential performance measures of the system. The sensitivity analysis results contribute to a better understanding of the studied queuing system and provide valuable insights for designing and optimizing such service systems. All numerical results are conducted on a Windows computer equipped with an 8th Gen Intel(R) Core (TM) i5-8250U CPU @ 1.80 GHz using the software MATLAB (R2018b).

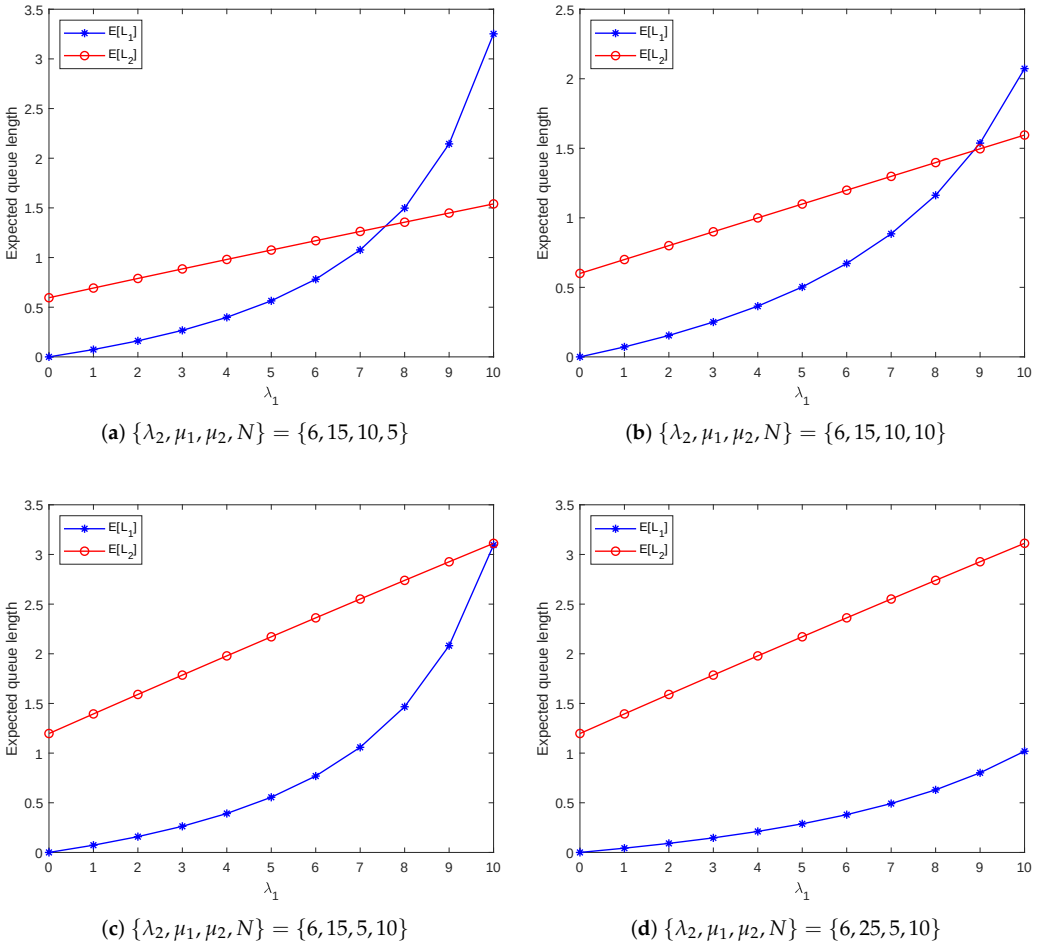
6.1. Sensitivity Analysis of System Parameters on Expected Queue Length

In Figure 4, the variation of the expected queue lengths $E[L_1]$ and $E[L_2]$ with respect to the arrival rate λ_1 is depicted. The graph illustrates four different parameter sets, showing that as λ_1 increases from 0 to 10, both $E[L_1]$ and $E[L_2]$ correspondingly increase. This indicates that a higher arrival rate λ_1 leads to longer queue lengths as the system accumulates more ordinary customers. It is worth noting that as λ_1 continues to rise, $E[L_2]$ increases proportionally, while $E[L_1]$ displays an exponential growth trend. This is because the second stage has finite capacity, which causes congestion in the first stage when the arrival rate λ_1 becomes larger.

Figure 5 examines the changes in $E[L_1]$ and $E[L_2]$ as functions of the arrival rate λ_2. With four different parameter sets, the figure demonstrates that when λ_2 increases from 0 to 10, both $E[L_1]$ and $E[L_2]$ also increase. This increase suggests that when more priority customers enter the second stage, there is a higher chance of the first stage being blocked. As a result, ordinary customers in the first stage may experience congestion.

Figure 5a,b show that when other parameters are fixed, $E[L_1]$ increases rapidly with an increase in λ_2 while $E[L_2]$ increases slowly when $N = 5$. In contrast, with $N = 10$, the trend reverses. This occurs because a larger N enhances the likelihood of open slots in the second stage, reducing the chance that the first stage becomes congested, thus allowing more ordinary and priority customers to advance to the second stage. Additionally, a comparison between Figure 5a,c reveals that at a constant λ_2, a higher service rate μ_1 decreases $E[L_1]$ but elevates $E[L_2]$. This is due to the increased flow of ordinary customers transitioning from

the first to the second stage as μ_1 rises. Lastly, the outcomes in Figure 5c,d demonstrate that when all other variables are constant, the values of $E[L_1]$ and $E[L_2]$ are lower at $\mu_2 = 5$ than at $\mu_2 = 3$. This is attributed to the fact that a lower μ_2 extends service times at the second stage, leading to a buildup of customers, which in turn increases the likelihood of blockages at the first stage.

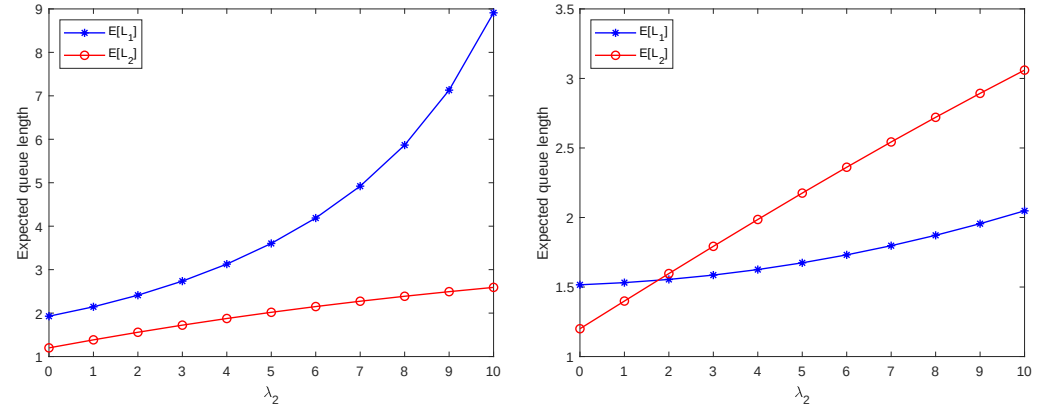


(**a**) $\{\lambda_2, \mu_1, \mu_2, N\} = \{6, 15, 10, 5\}$ (**b**) $\{\lambda_2, \mu_1, \mu_2, N\} = \{6, 15, 10, 10\}$

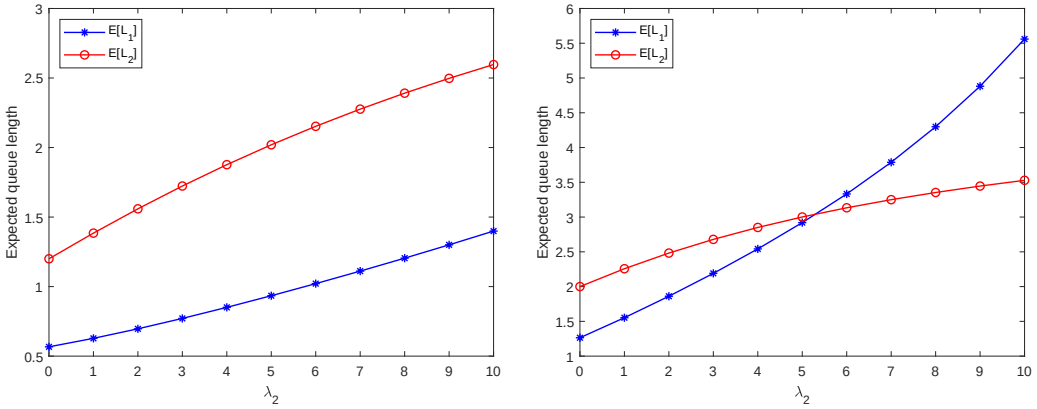
(**c**) $\{\lambda_2, \mu_1, \mu_2, N\} = \{6, 15, 5, 10\}$ (**d**) $\{\lambda_2, \mu_1, \mu_2, N\} = \{6, 25, 5, 10\}$

Figure 4. Expected queue length versus the arrival rate λ_1.

Figures 6 and 7 illustrate the impact of service rates μ_1 and μ_2 on the mean queue lengths $E[L_1]$ and $E[L_2]$. From Figure 6, we can see that as μ_1 increases, $E[L_1]$ initially decreases rapidly and then stabilizes. However, $E[L_2]$ does not show any significant change. This implies that increasing μ_1 below a certain threshold can significantly reduce the mean queue length in the first stage. However, if μ_1 is already above the threshold, further increase in the service rate will have a limited effect in reducing the mean queue length of the first stage. Figure 7 shows that $E[L_1]$ and $E[L_2]$ both gradually decrease with the increase of μ_2 and then tend to a constant value. When the service rate μ_2 is low, the second stage becomes a bottleneck for the system, resulting in queue backlog. At this point, increasing μ_2 appropriately can significantly reduce traffic intensity, decrease queue backlog, and significantly shorten the mean queue length. However, when μ_2 reaches a certain value, further increase μ_2 will not significantly change the queue length because customers in the system have already been served fast enough, and the additional service capacity has not been fully utilized.

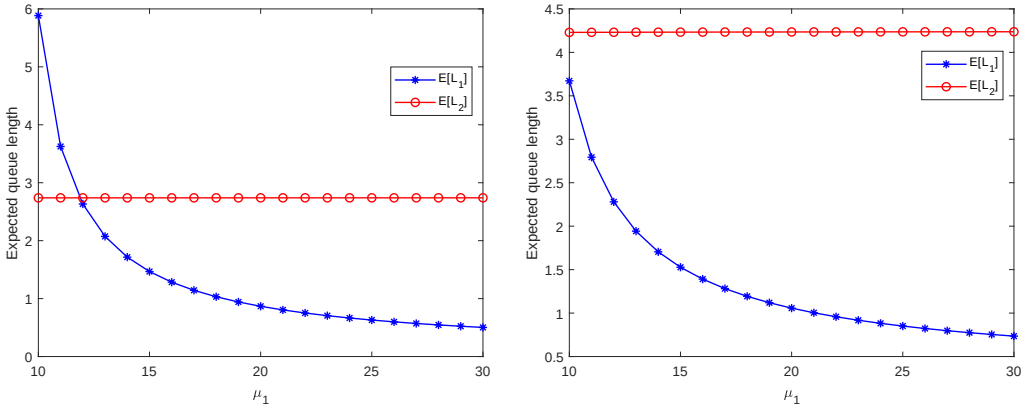


(**a**) $\{\lambda_1, \mu_1, \mu_2, N\} = \{6, 10, 5, 5\}$ (**b**) $\{\lambda_1, \mu_1, \mu_2, N\} = \{6, 10, 5, 10\}$



(**c**) $\{\lambda_1, \mu_1, \mu_2, N\} = \{6, 20, 5, 5\}$ (**d**) $\{\lambda_1, \mu_1, \mu_2, N\} = \{6, 20, 3, 5\}$

Figure 5. Expected queue length versus the arrival rate λ_2.

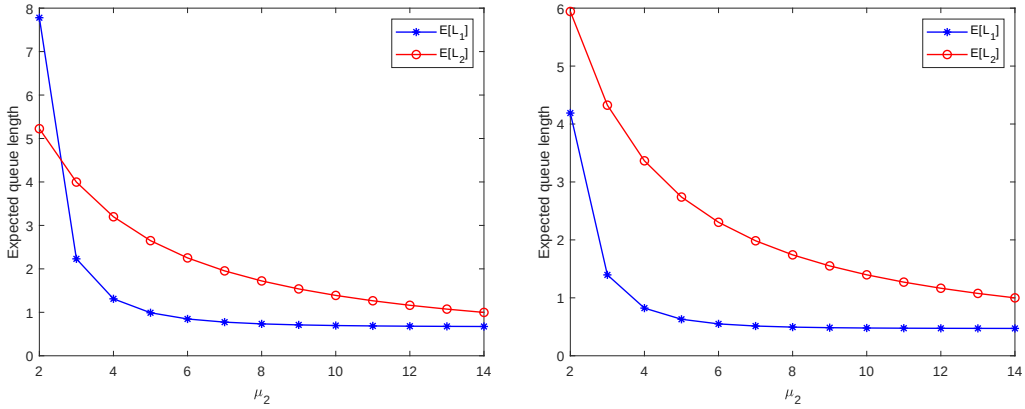


(**a**) $\{\lambda_1, \lambda_2, \mu_2, N\} = \{8, 6, 5, 10\}$ (**b**) $\{\lambda_1, \lambda_2, \mu_2, N\} = \{6, 8, 3, 10\}$

Figure 6. Expected queue length versus the service rate μ_1.

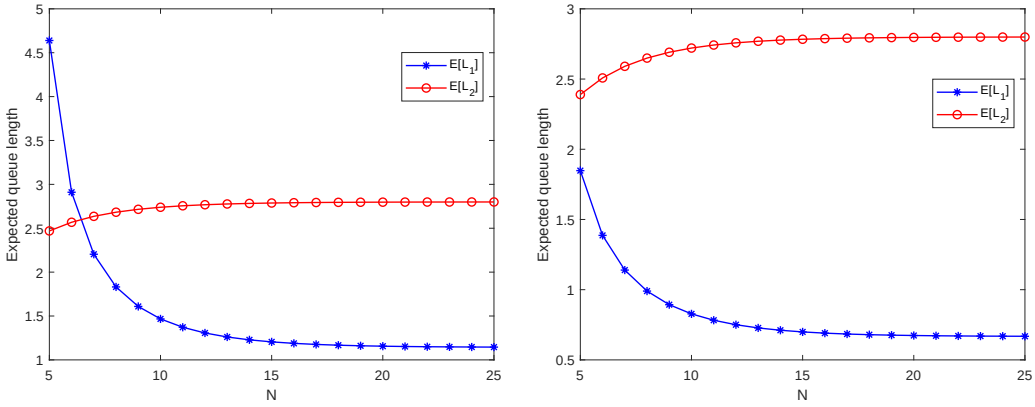
According to Figure 8, it is evident that when $N < 15$, $E[L_1]$ decreases drastically as N increases. However, when $N = 15$, $E[L_1]$ gradually stabilizes with the increase of N. This phenomenon arises due to the limited processing capacity of the second stage when N is small, leading to delays in customers from the first stage entering the second stage, which causes a backlog at stage one. As N increases, the second stage can accommodate more customers, facilitating a quicker transition for customers from the first stage to the second stage. Nevertheless, once N reaches a point where the second stage is no longer a bottleneck, the mean queue length in the first stage is no longer restricted by N. As for

$E[L_2]$, we can see that its change is insignificant and remains relatively stable as N increases. This indicates that the queue length in the second stage is not significantly affected by the threshold N. The reason may be the limitation of the service rate of μ_2. Even if more customers can be accommodated, the service capabilities may not match, so the queue length will not change significantly.



(**a**) $\{\lambda_1, \lambda_2, \mu_1, N\} = \{8, 6, 25, 10\}$ (**b**) $\{\lambda_1, \lambda_2, \mu_1, N\} = \{6, 8, 15, 8\}$

Figure 7. Expected queue length versus the service rate μ_2.



(**a**) $\{\lambda_1, \lambda_2, \mu_1, \mu_2\} = \{8, 6, 15, 5\}$ (**b**) $\{\lambda_1, \lambda_2, \mu_1, \mu_2\} = \{6, 8, 15, 5\}$

Figure 8. Expected queue length versus the threshold queue size of the second stage N.

6.2. Sensitivity Analysis of System Parameters on Expected Sojourn Time of an Ordinary Customer

Figures 9–12 show the expected sojourn time of ordinary customers in the system with respect to parameters λ_2, μ_1, μ_2, and N as the arrival rate λ_1 increases from 0 to 10. These figures show that the expected sojourn time of ordinary customers increases with the arrival rate λ_1 increase. Because more ordinary customers reach the first stage, newly arrived ordinary customers need to wait longer to receive the first stage of service, therefore increasing their overall sojourn time in the system.

According to Figure 9, when λ_1 is small (i.e., the arrival rate in the first stage is low), there are not many people in the system, which means the system is not close to saturation. As a result, the expected sojourn time of ordinary customers is relatively short. In such a scenario, even if more priority customers enter the second stage, the system still has sufficient service capability to serve customers, so there is no significant change in sojourn time. However, when λ_1 is large, the system approaches or reaches saturation. As λ_1 increases, the second stage becomes busier and may become a bottleneck for the system. This can cause ordinary customers to be blocked in the first stage, leading to an increase in wait time. Consequently, the overall sojourn time in the system increases.

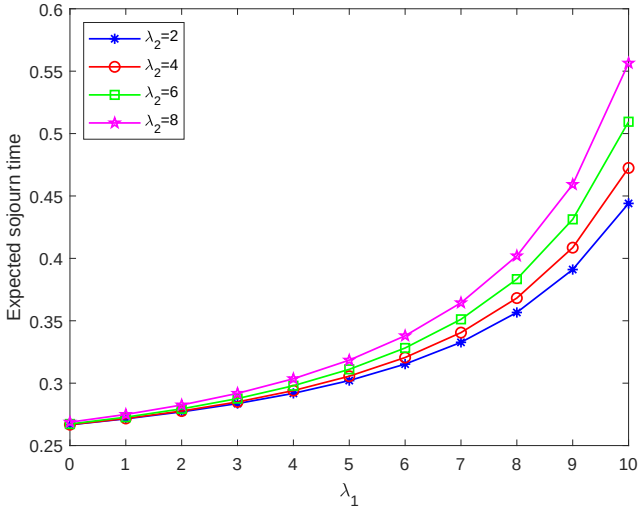


Figure 9. Expected sojourn time of ordinary customer in the system versus the arrival rate λ_2.

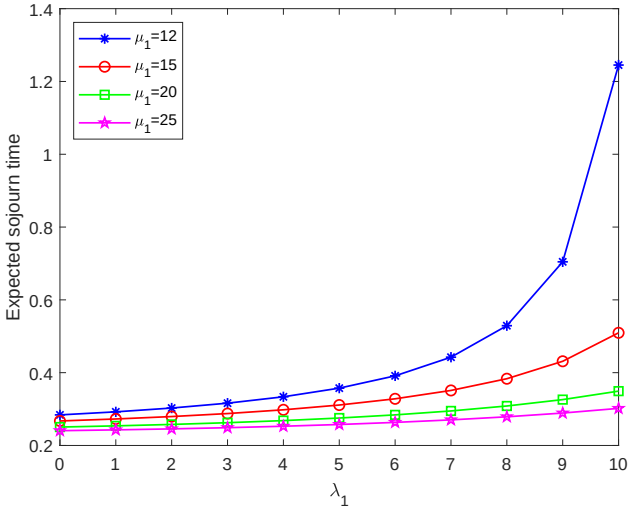


Figure 10. Expected sojourn time of ordinary customer in the system versus the service rate μ_1.

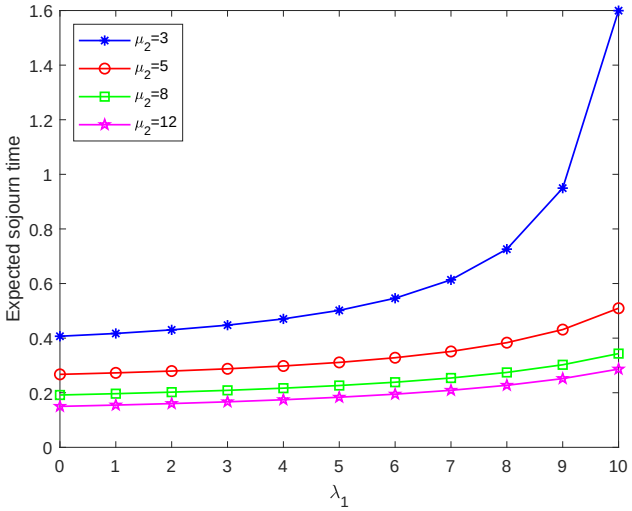


Figure 11. Expected sojourn time of ordinary customer in the system versus the service rate μ_2.

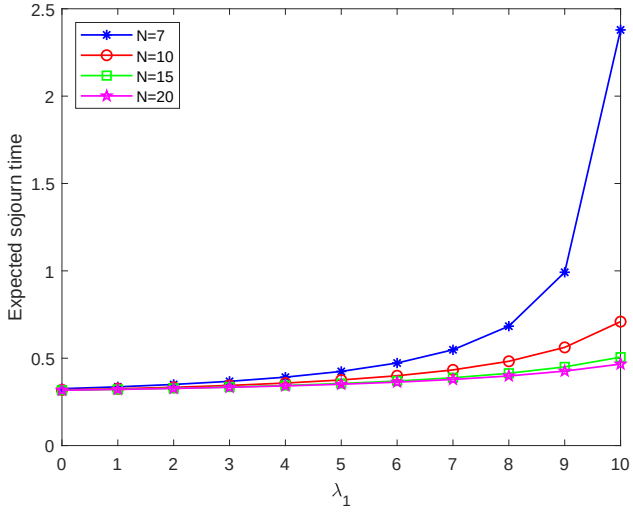


Figure 12. Expected sojourn time of ordinary customer in the system versus the threshold queue size of the second stage N.

As shown in Figures 10 and 11, for a given λ_1, the expected sojourn time of ordinary customers decreases with the increase of μ_1 and μ_2. This is because an increase in service rates leads to faster service for ordinary customers, thus reducing their sojourn time in the system. When λ_1 is small (such as $\lambda_1 < 4$), the difference in expected sojourn time for different μ_1 values is insignificant, and the expected sojourn time for different μ_2 values maintain low and similar values. However, when λ_1 is high (close to 10), the values of μ_1 and μ_2 significantly impact the expected sojourn time. In particular, when $\mu_1 = 12$ or $\mu_2 = 3$, the expected sojourn time increases sharply, indicating that a lower service rate will significantly reduce system efficiency under a high arrival rate. For higher service rates (such as $\mu_1 = 25$ or $\mu_2 = 12$), the expected sojourn time remains low even at a higher arrival rate. That is to say, improving service efficiency is an effective strategy to reduce customer sojourn time and improve overall service quality, especially when the arrival rate is large.

Figure 12 shows that when λ_1 is small, there is not much difference in the mean sojourn time between different N values. This indicates that the capacity threshold in the second stage does not significantly impact the mean sojourn time when ordinary customers arrive less frequently. When λ_1 is large, the expected sojourn time decreases with the increase of N, especially when N is large ($N = 20$). This means that when the system approaches saturation, a small N limits the ability of the second stage to process customers. By increasing the capacity threshold of the second stage, the pressure on the first stage can be reduced, and the overall system's smoothness can be improved. Therefore, setting a reasonable capacity threshold for the second stage is crucial for maintaining system efficiency, especially with a large arrival rate.

6.3. Sensitivity Analysis of System Parameters on Performance Measures

Tables 1–4 provide insights into how varying parameters λ_1, λ_2, μ_1, μ_2, and N on various system performance measures, including the probability that the first stage is empty P_{e1}, the probability that the second stage is empty P_{e2}, the probability that the system is empty P_e, the probability of the first stage being busy P_{b1}, the probability of the second stage being busy P_{b2}, and the probability of the second stage running at full load P_f.

As shown in Tables 1 and 2, note that as the values of λ_1 and λ_2 increase, the probabilities P_{e1}, P_{e2}, and P_e all decrease, which aligns with expectations. An increase in λ_1 leads to an increase in P_{b1}, P_{b2}, and P_f. However, when λ_2 increases, only P_{b2} and P_f increase, while P_{b1} remains constant because it is independent of the arrival rate of priority customers and is only determined by the parameters of the first stage.

Table 1. Performance measures versus λ_1 with $\lambda_2 = 6, \mu_1 = 15, \mu_2 = 5, N = 10$.

λ_1	P_{e1}	P_{e2}	P_e	P_{b1}	P_{b2}	P_f
1	0.9323	0.4164	0.3885	0.0667	0.5836	0.0046
2	0.8637	0.3840	0.3322	0.1333	0.6160	0.0078
3	0.7939	0.3559	0.2835	0.2000	0.6441	0.0122
4	0.7229	0.3313	0.2410	0.2667	0.6687	0.0176
5	0.6504	0.3096	0.2033	0.3333	0.6904	0.0243
6	0.5766	0.2901	0.1696	0.4000	0.7099	0.0320
7	0.5013	0.2726	0.1392	0.4667	0.7274	0.0408
8	0.4245	0.2568	0.1117	0.5333	0.7432	0.0507
9	0.3463	0.2423	0.0866	0.6000	0.7577	0.0616

Table 2. Performance measures versus λ_2 with $\lambda_1 = 8, \mu_1 = 15, \mu_2 = 5, N = 10$.

λ_2	P_{e1}	P_{e2}	P_e	P_{b1}	P_{b2}	P_f
1	0.4563	0.3545	0.1630	0.5333	0.6455	0.0125
2	0.4517	0.3299	0.1506	0.5333	0.6701	0.0182
3	0.4461	0.3083	0.1394	0.5333	0.6917	0.0249
4	0.4397	0.2892	0.1293	0.5333	0.7108	0.0326
5	0.4324	0.2721	0.1201	0.5333	0.7279	0.0413
6	0.4245	0.2568	0.1117	0.5333	0.7432	0.0507
7	0.4159	0.2430	0.1040	0.5333	0.7570	0.0608
8	0.4069	0.2305	0.0969	0.5333	0.7695	0.0715
9	0.3973	0.2191	0.0903	0.5333	0.7809	0.0827

Table 3. Performance measures versus μ_1 and N with $\lambda_1 = 8, \lambda_2 = 6, \mu_2 = 7$.

N	μ_1	P_{e1}	P_{e2}	P_e	P_{b1}	P_{b2}	P_f
5	10	0.0493	0.3115	0.0164	0.8000	0.6885	0.1545
	15	0.3388	0.3105	0.1129	0.5333	0.6895	0.1537
	20	0.4836	0.3098	0.1612	0.4000	0.6902	0.1532
	25	0.5705	0.3093	0.1902	0.3200	0.6907	0.1529
	30	0.6284	0.3089	0.2095	0.2667	0.6911	0.1526
10	10	0.1836	0.3310	0.0612	0.8000	0.6690	0.0179
	15	0.4526	0.3308	0.1509	0.5333	0.6692	0.0179
	20	0.5871	0.3307	0.1957	0.4000	0.6693	0.0179
	25	0.6678	0.3307	0.2226	0.3200	0.6693	0.0178
	30	0.7216	0.3306	0.2405	0.2667	0.6694	0.0178
15	10	0.1979	0.3330	0.0660	0.8000	0.6670	0.0023
	15	0.4649	0.3330	0.1550	0.5333	0.6670	0.0023
	20	0.5983	0.3330	0.1994	0.4000	0.6670	0.0023
	25	0.6784	0.3330	0.2261	0.3200	0.6670	0.0023
	30	0.7318	0.3330	0.2439	0.2667	0.6670	0.0023
20	10	0.1997	0.3333	0.0666	0.8000	0.6667	0.0003
	15	0.4664	0.3333	0.1555	0.5333	0.6667	0.0003
	20	0.5998	0.3333	0.1999	0.4000	0.6667	0.0003
	25	0.6798	0.3333	0.2266	0.3200	0.6667	0.0003
	30	0.7331	0.3333	0.2444	0.2667	0.6667	0.0003

Table 4. Performance measures versus μ_2 and N with $\lambda_1 = 8, \lambda_2 = 6, \mu_1 = 20$.

N	μ_2	P_{e1}	P_{e2}	P_e	P_{b1}	P_{b2}	P_f
5	3	0.0581	0.1105	0.0103	0.4000	0.8895	0.5535
	5	0.3698	0.2241	0.0973	0.4000	0.7759	0.2744
	8	0.5137	0.3452	0.1868	0.4000	0.6548	0.1181
	10	0.5495	0.4050	0.2290	0.4000	0.5950	0.0735
	15	0.5832	0.5130	0.3017	0.4000	0.4870	0.0272
10	3	0.4521	0.1585	0.0798	0.4000	0.8415	0.1714
	5	0.5607	0.2565	0.1476	0.4000	0.7435	0.0505
	8	0.5922	0.3620	0.2153	0.4000	0.6380	0.0111
	10	0.5969	0.4160	0.2487	0.4000	0.5840	0.0046
	15	0.5996	0.5171	0.3101	0.4000	0.4829	0.0007
15	3	0.5504	0.1704	0.0971	0.4000	0.8296	0.0589
	5	0.5920	0.2618	0.1558	0.4000	0.7382	0.0104
	8	0.5992	0.3635	0.2179	0.4000	0.6365	0.0011
	10	0.5998	0.4166	0.2499	0.4000	0.5834	0.0003
	15	0.6000	0.5172	0.3103	0.4000	0.4828	0.0000
20	3	0.5821	0.1743	0.1027	0.4000	0.8257	0.0213
	5	0.5983	0.2629	0.1574	0.4000	0.7371	0.0022
	8	0.5999	0.3636	0.2182	0.4000	0.6364	0.0001
	10	0.6000	0.4167	0.2500	0.4000	0.5833	0.0000
	15	0.6000	0.5172	0.3103	0.4000	0.4828	0.0000

Table 3 shows that as μ_1 increases, P_{e1} increases while P_{b1} decreases. The changes in P_{e2} and P_e indicate that although the service rate in the first stage improves its own performance, it has little impact on the idle state of the second stage and the entire system. Table 4 presents that when μ_2 increases, P_{e2} significantly increases, while P_{b2} and P_f decrease. Therefore, increasing μ_1 mainly improves the congestion situation of the first stage, while increasing μ_2 not only reduces the pressure of the second stage but also reduces the congestion of the whole system and the customer's sojourn time in the system.

In addition, from Table 3, it also can be observed that as the value of N increases, P_{e1} usually increases, while P_{e2} slightly decreases. This implies that the second stage of the system can accommodate more customers, which ultimately relieves the pressure on the first stage. However, P_e and P_{b1} do not show a consistent trend because they are more affected by μ_1 and μ_2. The probability of full-load operation in the second stage, P_f, has hardly changed, which may be because even when μ_2 is relatively small, the second stage will still quickly reach full load, even if N increases.

As shown in Table 4, with the increase of N, P_{e1} increases while P_{e2} decreases, which is consistent with the trend in Table 3. However, when μ_2 is large, P_f significantly decreases, particularly when N is small. In summary, the impact of N on system performance indicates that increasing the second stage's queue capacity can alleviate pressure on the first stage by allowing more customers into the second stage's queue. Nonetheless, this effect also depends on the service rate μ_2. With a lower μ_2, increasing N might not substantially reduce the busyness of the second stage, as service efficiency is the limiting factor. Conversely, increasing N can effectively reduce the probabilities of busyness and full-load operation when μ_2 is sufficiently high, therefore optimizing system performance.

7. Conclusions

We analyze a two-stage tandem queuing system consisting of both ordinary and priority customers. Ordinary customers first receive service at stage one before moving on to stage two for clearing service. Priority customers bypass the first stage, directly entering the second stage for their service. We first model this system as a two-dimensional Markov chain to study the system stability condition. Subsequently, we employ the matrix-analytic method alongside spectral expansion to derive the system's stationary distribution.

We also offer analytical formulas for key characteristics, such as the expected sojourn time of ordinary customers, expected queue length, and probability of the second stage running full load. Finally, we examine the effects of various system parameters through numerical examples, providing insights that could guide the design of similar two-stage service systems.

Author Contributions: Conceptualization, J.X. and L.L.; methodology, J.X. and L.L.; software, J.X.; validation, J.X.; formal analysis, J.X.; writing—original draft preparation, J.X.; writing—review and editing, J.X. and L.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the National Natural Science Foundation of China (Grant No. 61773014).

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Balsamo, S.; Persone, V.D.N.; Inverardi, P. A review on queuing network models with finite capacity queues for software architectures performance prediction. *Perform. Eval.* **2003**, *51*, 269–288. [CrossRef]
2. Balsamo, S. Queueing networks with blocking: Analysis, solution algorithms and properties. In *Network Performance Engineering: A Handbook on Convergent Multi-Service Networks and Next Generation Internet*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 233–257.
3. Wang, J.; Abouee-Mehrizi, H.; Baron, O.; Berman, O. Tandem queues with impatient customers. *Perform. Eval.* **2019**, *135*, 102011. [CrossRef]
4. Do, T.V. A closed-form solution for a tollbooth tandem queue with two heterogeneous servers and exponential service times. *Eur. J. Oper. Res.* **2015**, *247*, 672–675. [CrossRef]
5. Jackson, J.R. Networks of waiting lines. *Oper. Res.* **1957**, *5*, 518–521. [CrossRef]
6. VanOyen, M.P.; Teneketzis, D. Optimal stochastic scheduling of forest networks with switching penalties. *Adva. Appl. Probab.* **1994**, *26*, 474–497. [CrossRef]
7. Katayama, T. Performance analysis and optimization of a cyclic service tandem queueing system with multi-class customers. *Comput. Math. Appl.* **1992**, *4*, 25–33. [CrossRef]
8. Chang, K.-H.; Chen, W.-F. Admission control policies for two-stage tandem queues with no waiting spaces. *Comput. Oper. Res.* **2003**, *30*, 589–601. [CrossRef]
9. Neuts, M.F. Two queues in series with a finite, intermediate waiting room. *J. Appl. Probab.* **1968**, *5*, 123–142. [CrossRef]
10. Neuts, M. *Matrix-Geometric Solutions in Stochastic Models*; The Johns Hopkins University Press: Baltimore, MD, USA, 1981.
11. Yang, W.-S.; Kim, T.; Park, H.; Lim, D. Analysis of a two-stage queue with a single server and N-policy. *Am. J. Math. Manag. Sci.* **2016**, *35*, 261–270. [CrossRef]
12. Nazarov, A.; Phung-Duc, T.; Paul, S.; Morozova, M. Scaling limits of a tandem queue with two infinite orbits. *Mathematics* **2023**, *11*, 2454. [CrossRef]
13. Dudin, S.A.; Dudina, O.S.; Dudin, A.N. Analysis of tandem queue with multi-server stages and group service at the second stage. *Axioms* **2024**, *13*, 214. [CrossRef]
14. Choi, B.D.; Chang, Y. Single server retrial queues with priority calls. *Math. Comput. Modell.* **1999**, *30*, 27–32. [CrossRef]
15. Gómez-Corral, A. Analysis of a single-server retrial queue with quasi-random input and non-preemptive priority. *Comput. Math. Appl.* **2020**, *43*, 767–782. [CrossRef]
16. Walraevens, J.; Steyaert, B.; Bruneel, H. A preemptive repeat priority queue with resampling: Performance analysis. *Ann. Oper. Res.* **2006**, *146*, 189–202. [CrossRef]
17. Artalejo, J.R.; Dudin, A.N.; Klimenok, V.I. Stationary analysis of a retrial queue with preemptive repeated attempts. *Oper. Res. Lett.* **2001**, *28*, 173–180. [CrossRef]
18. Kim, C.; Klimenok, V.I.; Dudin, A.N. Priority tandem queueing system with retrials and reservation of channels as a model of call center. *Comput. Ind. Eng.* **2016**, *96*, 61–71. [CrossRef]
19. Liu, B.; Zhao, Y.Q. Tail asymptotics for the $M_1, M_2/G_1, G_2/1$ retrial queue with non-preemptive priority. *Queueing Syst.* **2020**, *96*, 169–199. [CrossRef]
20. Lee, S.; Dudin, S.; Dudina, O.; Kim, C.; Klimenok, V. A priority queue with many customer types, correlated arrivals and changing priorities. *Mathematics* **2020**, *8*, 1292. [CrossRef]
21. Kim, C.; Dudin, S. Priority tandem queueing model with admission control. *Comput. Ind. Eng.* **2011**, *61*, 131–140. [CrossRef]
22. Atencia, I. A Geo/G/1 retrial queueing system with priority services. *Eur. J. Oper. Res.* **2017**, *256*, 178–186. [CrossRef]
23. Xie, J.; Zhu, T.; Chao, A.K.; Wang, S. Performance analysis of service systems with priority upgrades. *Ann. Oper. Res.* **2017**, *253*, 683–705. [CrossRef]

24. Xu, J.; Liu, L.; Wu, K. Analysis of a retrial queueing system with priority service and modified multiple vacations. *Comm. Stat. Theory Methods* **2023**, *52*, 6207–6231. [CrossRef]
25. Chamberlain, J.; Starobinski, D. Social welfare and price of anarchy in preemptive priority queues. *Oper. Res. Lett.* **2020**, *48*, 530–533. [CrossRef]
26. Mitrani, I.; Chakka, R. Spectral expansion solution for a class of Markov models: application and comparison with the matrix-geometric method. *Perform. Eval.* **1995**, *23*, 241–260. [CrossRef]
27. Haverkort, B.R.; Ost, A. Steady-state analysis of infinite stochastic Petri nets: Comparing the spectral expansion and the matrix-geometric method. In Proceedings of the Seventh International Workshop on Petri Nets and Performance Models, Saint Malo, France, 3–6 June 1997; pp. 36–45.