

Article

Customer Equilibrium and Optimal Strategies in a Batch-Service Queueing System with a Proportional Pricing Coefficient

Pingping Wang ¹, Wenjing Peng ¹, Yuquan Yao ¹, Yingling Chen ¹ and Zhen Wang ^{1,2,*} 

¹ Guangxi Colleges and Universities Key Laboratory of Data Analysis and Computation, School of Mathematics and Computing Science, Guilin University of Electronic Technology, Guilin 541004, China; wangpp0219@163.com (P.W.)

² Center for Applied Mathematics of Guangxi (GUET), Guilin 541004, China

* Correspondence: wang686zhen@guet.edu.cn

Abstract

This work analyzes an observable batch service queueing system with a proportional pricing coefficient and multiple access points. Based on whether the system is idle upon arrival, we derive the equilibrium threshold strategy for customers to either wait for batch formation or take service alone. When the server is idle, we derive the optimal waiting rate for customers in the queue. When it is occupied, we analyze the equilibrium threshold strategy and find the threshold value via a binary search algorithm. Furthermore, we quantify the system's key performance measures for subsequent numerical experiments, establish a bi-level optimization model to maximize social welfare, and determine the optimal proportional pricing coefficient via a grid search. Finally, through numerical experiments, we conduct sensitivity analyses of the optimal proportional pricing coefficient, examine the impact of parameters such as the arrival rate and the number of access points on performance measures, and demonstrate the advantages of the strategic batch mechanism over the fixed batch mode.

Keywords: customer strategies; performance measures; proportional pricing coefficient; batch service queueing system; access point

MSC: 60K20; 60K25; 60K30



Academic Editor: Mark Kelbert

Received: 8 May 2026

Revised: 9 July 2026

Accepted: 14 July 2026

Published: 16 July 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The balance between operational revenue and customer experience is a key consideration for sharing economy platform managers. Motivated by these two key factors, this work constructs an observable M/M/S queueing system with strategic batching and dynamic pricing. The proposed model has potential applications in the scheduling scenario of two-person pedal boats in scenic areas. At many scenic water spots, tourists need to take boats from the dock to reach the attractions. The operator provides multiple pedal boats for simultaneous service. Each pedal boat can accommodate up to two passengers, and the service time per trip is stochastic, regardless of the number of tourists served. Considering the operating costs and resource utilization of pedal boats, the operator has introduced a dynamic pricing mechanism: If a tourist chooses to charter a boat, he/she will have to pay the full fare, and if he/she chooses to wait to share the boat with others, they can share the cost proportionally. As tourists can observe the situation at the dock in real time

through the mobile APP, rational tourists will consider whether to charter a boat alone or share one based on the waiting cost and the price discount. In this scenario, tourists, pedal boats, waiting for boat-sharing decisions, dynamically adjusted pricing, and operating boat numbers correspond to the customer, server, batch strategy, pricing strategy, and control access point strategy in queuing-theory terminology, respectively.

As another application of our batch model, we consider a two-person rickshaw service in a scenic area. Considering the physical exertion and operating costs for rickshaw drivers, the scenic area provides a sightseeing rickshaw with a fixed capacity of 2 passengers. To encourage tourists to carpool and improve resource utilization, managers have introduced a dynamic pricing mechanism. Tourists can either pay the full price to rent a rickshaw and leave immediately, or wait and share a ride with another tourist to split the cost. As tourists can observe the queue length, tourists will consider whether to charter or to carpool based on waiting time and the price discount. During peak tourism periods, scenic area managers will dynamically adjust the number of rickshaws in operation to balance service supply and tourist demand. In this scenario, tourists, rickshaws, carpooling decisions, cost-sharing mechanisms, and the number of operating vehicles correspond to the customer, server, batch strategy, pricing strategy, and capacity control strategy in queuing theory terminology, respectively.

The balancing decision of tourists, based on waiting time and the price discount in the above scenario, essentially falls under the strategic customer game problem in queuing economics. Tracing back to the research origins of this field, Naor's groundbreaking work [1] introduced customer strategic behavior in observable queuing systems for the first time, revealing the deviation between Nash equilibrium and social optimality caused by customers pursuing individual utility maximization. Subsequently, Edelson and Hildebrand [2] further extended this analytical framework to the study of stopping strategies in unobservable systems, enriching the theoretical foundation of queuing games. These two scenarios constitute the two cornerstones of queuing game research and have given rise to a large number of studies under different settings, such as nonexponential service distributions [3,4], reneging behavior [5,6], feedback queues [7,8], and preemptive queues [9,10].

However, in traditional $M/M/1$ single-server queuing systems, adding customers only increases system congestion. Once the number of customers increases, newly arrived customers become less willing to join. When we turn to batch service queuing systems, this common sense no longer holds. On the one hand, the increase in complete batches in the system prompts customers to choose to balk. On the other hand, the increase in the number of customers in the current batches has prompted more customers to join. This phenomenon leads to batch service systems of more complex game characteristics. In recent years, an increasing number of scholars have begun to study batch service queuing systems. Bountali and Economou [11] studied the equilibrium joining strategies of customers in a fixed-batch service queue under both observable and unobservable cases, finding that batch service generates both positive and negative externalities and may lead to multiple equilibria. Nguyen and Phung-Duc [12] further incorporated the decisions of "whether to wait for batch formation" and "whether to join the queue" into a unified framework, deriving the optimal waiting time and threshold-type joining strategy for customers in an observable batch-service queue. Nakamura and Phung-Duc [13] studied a variable batch-service queueing system in which customers can strategically choose the batch size. In their model, customers in the same batch share the service fee equally, and, based on this assumption, they analyzed equilibrium joining strategies and system performance. Sun et al. [14] studied Markovian queues with a batch-limited service rule and multiple vacations, finding that there may be multiple equilibria, but not all of them are stable. Moreover, in the fully unobservable case, customers' equilibrium behavior

is inconsistent with their socially optimal behavior, yet an optimal batch size exists that maximizes social welfare. Liu et al. [15] investigated an unreliable M/M/1+1 retrial queue model featuring batch service provided by a standby server, analyzed the equilibrium strategies for two types of parallel customers, and the system's average social benefits under different information levels. Hanukov et al. [16] examined the group-merging Israeli queue, in which customers can merge two adjacent waiting customers and their respective groups into a single group for batch service, derived customer-joining thresholds, and showed that the queue can remain unbounded regardless of customers' strategic behavior.

In the batch service scenario of the sharing economy, the cost-sharing mechanism is the primary incentive influencing the decision to form batches. Traditional batch-service models typically assume that customers share the service fee equally, i.e., each customer pays P/K , where P is the total service fee and K is the number of customers. However, this equal-sharing assumption overlooks the asymmetry in waiting time between the batch initiator and the subsequent joiner. To capture such differentiated incentive designs, some studies have introduced the concept of a proportional pricing coefficient or a revenue-sharing ratio. Fielbaum et al. [17] investigated cost-sharing agreements in ride-sharing settings that align system optimality with user equilibrium. Wang et al. [18] studied pre-sale and upgrade issues in priority queues under different pricing schemes. Nakamura and Phung-Duc [19] proposed a dynamic pricing control model for batch service systems, in which the fee is adjusted based on customers' waiting times to achieve within-batch fairness. Bergquist and Elmachtoub [20] provided performance guarantees for static pricing in queueing systems with price-sensitive customers. Inspired by this, the present work introduces a proportional pricing coefficient α into the batch cost-sharing scheme, in which the initiator pays αP and the subsequent customer pays $(1 - \alpha)P$, thereby capturing the differential waiting-cost burden that naturally arises in batch-service systems.

When the differentiated cost-sharing mechanism mentioned above is introduced, customers' decision-making space also changes accordingly. For customers who choose batch formation, the system charges service fees based on the proportional pricing coefficient α . After observing the system states, each customer decides whether to serve separately, wait for a batch, or balk based on the reward-cost structure. Therefore, there are two questions that motivate us to study the strategies and control mechanisms of such systems: (1) How to export the balanced strategy of mixed services for this work? (2) How do the proportional pricing coefficient α and the number of access points affect the performance measures of the system? This work answers these two questions.

This work also provides some management insights for platform managers in the sharing economy. And the main contributions of this work are as follows.

1. This work introduces the proportional pricing coefficient and access points into an observable batch service queueing system, successfully characterizing the incentive difference between the initiator and the latecomer of a shared scenario due to asymmetric waiting time.
2. Derive and simulate the equilibrium strategy of customers. Design a binary search algorithm to solve threshold values that are difficult to parse. A general M/M/S formula is derived, and the steady-state probabilities are solved using the Gauss–Seidel iteration. Establish a dual-layer optimization model and use a grid search algorithm to search for maximal social welfare for a given λ .
3. Through extensive numerical experiments, we explore the impact and sensitivity of the proportional pricing coefficient and access points on customers' expected payoff, as well as their effects on customer behavior and system throughput.

The rest of this paper is organized as follows. In Section 2, we present the model description and parameter assumptions. In Section 3, we determine the optimal waiting rate and the customer equilibrium threshold strategy under the observable case. In Section 4,

we quantify the key performance measures. In Section 5, we conduct a series of numerical experiments to investigate the impact of key parameters on the key performance measures. Finally, in Section 6, we present conclusions and further discussions.

2. Model Description

The queuing model considered in this work is an observable M/M/S queuing system with strategic batching and dynamic pricing. The system is a single service facility consisting of S parallel access points that operate simultaneously (i.e., a service pool). Each access point can accommodate a batch of up to two customers at a time. Upon arrival, the current number of customers in the queue can be observed. Based on this, customers decide whether to join the system to maximize their expected payoff. If a customer decides to join when at least one access point is idle, they can choose to use the service immediately or wait for the next customer to share the access point. Otherwise, if all access points are occupied, arriving customers must decide whether to join the queue or leave the system; their decisions are irrevocable. To characterize cost-sharing behavior, this work introduces the proportional pricing coefficient α . If the customer uses the service alone, they will need to pay the full service fee; if two customers share an access point, the customer who enters the waiting area first will pay α times the service fee, and the subsequent customer will pay $(1 - \alpha)$ times the service fee. During peak hours, the system administrator will dynamically adjust the number of access points S and the proportional pricing coefficient α based on the number of customers in the system.

Suppose that each customer arrives at the system according to a Poisson process with parameter rate λ , and the arrival intervals are independent of each other. Denote the arrival interval of two adjacent customers as $T \sim \exp(\lambda)$. Assuming that the service time follows an exponential distribution with a rate of μ , and the service time is independent of the number of customers served, denote the personal service time of each customer as $T_s \sim \exp(\mu)$. The state of the system can be fully observed when the customer arrives. The system state is denoted as (m, j) , where $m \geq 0$ represents the total number of batches in the system, and j indicates the parity of the queue length, where $j = 0$ denoting an even queue length, and $j = 1$ denoting an odd queue length. Assuming that the service value of each customer is V , and the service price is P , then there must be $V > P > 0$. Set C_w as the waiting cost per unit time of the customer. Customers' decisions are made in two stages. First, decide whether to join the queue. If they become unmatched after joining, they need to decide whether to wait for the next customer to form a batch. If the customer uses the service without waiting, they need to pay P . Second, if the customer chooses to wait, they need to pay αP , and the other customer who forms a batch needs to pay $(1 - \alpha)P$. Once the customer decides to wait for the other customer to form a batch, he must register a maximum waiting time T_w , and pay a fee of C_r per unit of registered waiting time, where T_w following the exponential distribution with parameter ω , and is denoted as $T_w \sim \exp(\omega)$. The customer's net benefit, v_0 , is the difference between the value of the service and the actual service fee paid. Figure 1 depicts the operational mechanism of this work.

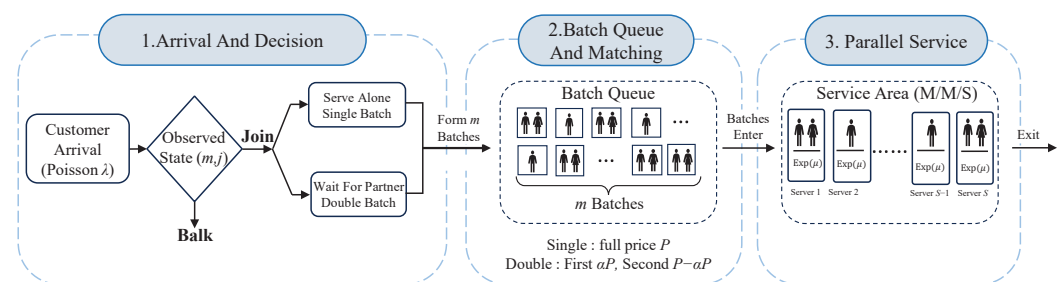


Figure 1. Operational mechanism of this queuing system.

The notations used in the work are shown in Table 1.

Table 1. Description of parameters and symbols in this work.

Notations	Explanations
λ	Customer arrival rate
T	Arrival time interval of customers
μ	Service rate
T_s	Customer service time
V	Customer service value
P	Total price of batch service
v_0	The customer's net benefit
C_w	Waiting cost per unit time
C_r	Booking cost per unit time
α	Proportional pricing coefficient
S	Number of access points
T_w	Maximum waiting time for customer registration
ω	Customer registration waiting rate
T_v	The actual waiting time for a tagged customer to form a complete batch
ω^*	Optimal waiting rate of the customer
C_m	The total number of rounds for the first m batches of customers
T_m	The total round time of the first m batches of customers
$W(m, 1)$	Customer's sojourn time in system state $(m, 1)$, $m \neq 0$
$W(m, 0)$	Customer's sojourn time in system state $(m, 0)$, $m \neq 0$
m_i^e	The optimal threshold when observing system state (m, i) , $m \neq 0$
Q	Average number of customers in the system
$U_1(m)$	Expected payoff of customers in system state $(m, 1)$, $m \neq 0$
$U_0(m)$	Expected payoff of customers in system state $(m, 0)$, $m \neq 0$
$p_{x,y}$	Steady state probability of state space (x, y)
ρ	Resource utilization
λ_e	Effective arrival rate
I_a	Total revenue generated by the system per unit time
P_b	Customer blocking probability
(x, y)	State space
(m, j)	Customer-observable system state: m is the total number of batches in the system, and j indicates the parity of the queue length
SW	Social welfare
$f(\omega)$	Expected payoff of customers in system state $(0, 0)$
$\{(X(t), Y(t))\}$	A continuous time Markov chain: $X(t)$ is the number of customers waiting in the queue at time t , and $Y(t)$ is the service mode at time t
$\mathcal{Y}_S$	The set of customers served
$\mathcal{B}$	Blocking state set
$\mathcal{Y}$	The set of possible numbers of customers in service, $0 \notin \mathcal{Y}$
M	The size of set $\mathcal{Y}$
α^*	The optimal proportional pricing coefficient
C_o	Operating cost per completed batch service
C_S	Fixed cost per access point per unit time

3. Analysis of Customer Behavior Strategies

This section will discuss the customer's equilibrium strategy in two scenarios, depending on whether the server is occupied when the customer arrives at the system, with α given.

3.1. Scenario 1: Arriving at the System When the System Is Idle ($m = 0$)

A tagged customer arrives at the system and finds the system is idle and an unmatched customer waiting, which is the state $(0, 1)$. The optimal strategy for tagged customers is to immediately form a batch with the other customer without waiting. The sojourn time of the

tagged customer is thus exactly the service time $\frac{1}{\mu}$, and the term $C_w \frac{1}{\mu}$ accounts for the service time cost during service. Therefore, the expected payoff is given by $V - (1 - \alpha)P - C_w \frac{1}{\mu}$.

A tagged customer arrives and finds the system is idle, with no one waiting in line, which is the state $(0, 0)$. If he/she joins, he/she becomes an unmatched customer, and the system state changes to $(0, 1)$. Now he/she faces a decision in the second stage: using the service alone or waiting for a batch. Assuming the customer decides to wait for the next customer to form a batch, he/she aims to maximize his/her expected payoff, defined as the service value minus the waiting cost and the service fee. Let T_v be the actual waiting time for a tagged customer to form a complete batch, then $T_v = \min\{T, T_w\}$, and $T_v \sim \exp(\lambda + \omega)$. If $T \leq T_w$, the two customers successfully form a batch, and the tagged customer pays αP , yielding a net payoff of $V - \alpha P$. If $T_w < T$, the tagged customer chooses the individual service, yielding a net payoff of $V - P$. Therefore, the expected net payoff function for the tagged customer is given by

$$\begin{aligned}\mathbb{E}[v_0] &= \Pr(T_w \geq T) \cdot (V - \alpha P) + \Pr(T_w < T) \cdot (V - P) \\ &= \frac{\omega}{\omega + \lambda}(V - P) + \frac{\lambda}{\omega + \lambda}(V - \alpha P).\end{aligned}\quad (1)$$

Whether or not the batch formation succeeds, the customer incurs the waiting time cost $C_w \cdot T_v$, the service time cost $C_w \cdot T_s$, and a reservation cost $C_r \cdot T_w$ that is proportional to the declared maximum waiting time. Therefore, the tagged customer's expected payoff is

$$\begin{aligned}f(\omega) &= \mathbb{E}[v_0 - C_w T_v - C_w T_s - C_r T_w] \\ &= \mathbb{E}[v_0] - C_w \min\{T, T_w\} - C_w \mathbb{E}[T_s] - C_r \mathbb{E}[T_w] \\ &= \frac{\omega}{\omega + \lambda}(V - P) + \frac{\lambda}{\omega + \lambda}(V - \alpha P) - \frac{C_w}{\omega + \lambda} - \frac{C_w}{\mu} - \frac{C_r}{\omega}. \\ &= V - P + \frac{\lambda}{\omega + \lambda}P(1 - \alpha) - C_w \left(\frac{1}{\omega + \lambda} + \frac{1}{\mu} \right) - C_r \frac{1}{\omega}.\end{aligned}\quad (2)$$

In this case, the optimal waiting strategy for customers is given by Theorem 1.

Theorem 1 (Optimal waiting strategy). *In an observable M/M/S queuing system, for a tagged customer with system state $(0, 0)$, it is necessary to decide whether to take the available server directly or wait for the next customer to form a batch-shared service. For a given α , the optimal waiting rate ω^* is determined by the following equation.*

$$\omega^* = \begin{cases} +\infty & \text{if } (1 - \alpha)P\lambda \leq C_w + C_r, \\ \frac{C_r\lambda + \lambda\sqrt{C_r(\lambda(1 - \alpha)P - C_w)}}{\lambda(1 - \alpha)P - C_w - C_r} & \text{if } (1 - \alpha)P\lambda > C_w + C_r. \end{cases}\quad (3)$$

Proof of Theorem 1. For a customer who observes system state $(0, 0)$, the expected payoff is given by Equation (2). To derive the maximum of $f(\omega)$, take its derivative with respect to ω . The first-order derivative is as follows.

$$f'(\omega) = \frac{(\lambda(\alpha - 1)P + C_w + C_r)\omega^2 + 2C_r\lambda\omega + C_r\lambda^2}{\omega^2(\omega + \lambda)^2}.\quad (4)$$

To determine the optimal ω , we consider the following two cases:

Case 1. If $H = \lambda(\alpha - 1)P + C_w + C_r \geq 0$, then $f'(\omega) > 0$ for all $\omega \in [0, +\infty)$, implying that $f(\omega)$ is monotonically increasing on $[0, +\infty)$. In this case, the expected payoff is maximized as $\omega = +\infty$; that is, the tagged customer's best response is to immediately take

the server without waiting. Consequently, rational customer behavior results in no batch service, so the equilibrium system is identical to Naor's model.

Case 2. If $H = \lambda(\alpha - 1)P + C_w + C_r < 0$, the numerator of the derivative $f'(\omega)$ is a quadratic function in ω . Let $y(\omega) = H\omega^2 + 2C_r\lambda\omega + C_r\lambda^2$, since $y(0) = C_r\lambda^2 > 0$ and $y(\omega) \rightarrow -\infty$ as $\omega \rightarrow +\infty$, the quadratic must have two real roots. The product of the two roots is $\frac{C_r\lambda^2}{H} < 0$, the roots have opposite signs, so the equation $f'(\omega) = 0$ has a unique positive root, given by

$$\omega^* = \frac{C_r\lambda + \lambda\sqrt{C_r(\lambda(1-\alpha)P - C_w)}}{\lambda(1-\alpha)P - C_w - C_r}. \quad (5)$$

When $\omega < \omega^*$, $f'(\omega) > 0$; when $\omega > \omega^*$, $f'(\omega) < 0$. Therefore, $f(\omega)$ attains its maximum at $\omega = \omega^*$. This is also the equilibrium strategy for a customer who observes an idle server and no waiting customers. This completes the proof. $\square$

3.2. Scenario 2: Arriving at the System When the Server Is Occupied ($m \neq 0$)

Now consider the case where the tagged customer arrives when the system is occupied (i.e., busy), and there are m batches of customers waiting ahead of the tagged customer. Based on the number of customers currently in line, the customer first decides whether to join the queue, and then considers the scenario discussed in Section 3.1.

The total number of rounds for the first m batches of customers is denoted by C_m . Since each round can serve S batches, we have $C_m = \lceil \frac{m}{S} \rceil$, where $\lceil z \rceil$ represents the smallest integer not less than z . Let T_m be the total round time for the first m batches of customers, and the service time for each round be $\frac{1}{\mu}$. Therefore, $T_m = \lceil \frac{m}{S} \rceil \frac{1}{\mu}$. T_v represents the actual waiting time for tagged customers to form a complete batch. In state $(m, 0)$, T_v is the actual waiting time that begins after the completion of service for the first m complete batches, i.e.,

$$T_v = \begin{cases} \min\{T - T_m, T_w\} & , T > T_m \\ 0 & , T \leq T_m \end{cases} \quad (6)$$

Based on the observed number of unmatched customers in the queue, the sizes of the access points, and the number of m batches, a tagged customer will face one of the following four scenarios.

Case 1. The system state is $(m, 1)$ and $m < S$. The tagged customer immediately forms the $(m + 1)$ th batch with the unmatched customers and is served in the same round as the first m batches. Thus, the actual waiting time for a customer to form a complete batch is 0, the maximum registered waiting time is 0, the sojourn time $W(m, 1)$ equals the total round time, $\mathbb{E}[W(m, 1)] = \frac{1}{\mu}$, and the service fee is $(1 - \alpha)P$. The customer's expected payoff $U_1(m)$ is

$$\begin{aligned} U_1(m) &= V - (1 - \alpha)P - C_w\mathbb{E}[W(m, 1)] \\ &= V - (1 - \alpha)P - C_w\frac{1}{\mu}. \end{aligned} \quad (7)$$

Case 2. The system state is $(m, 1)$ and $m \geq S$. The tagged customer immediately forms the $(m + 1)$ th batch with the unmatched customers, and the system provides service once every S batches. Thus, the actual waiting time for a customer to form a complete batch is 0, the maximum registered waiting time is 0, the sojourn time $W(m, 1)$ is the total round time, $\mathbb{E}[W(m, 1)] = \lceil \frac{m+1}{S} \rceil \frac{1}{\mu}$, and the service fee is $(1 - \alpha)P$. The customer's expected payoff $U_1(m)$ is

$$\begin{aligned} U_1(m) &= V - (1 - \alpha)P - C_w\mathbb{E}[W(m, 1)] \\ &= V - (1 - \alpha)P - C_w\lceil \frac{m+1}{S} \rceil \frac{1}{\mu}. \end{aligned} \quad (8)$$

Case 3. The system state is $(m, 0)$ and $m < S$. In this case, the customer must wait until service for the first m batches is completed, with a waiting time of T_m , where $T_m = (\lceil \frac{m}{S} \rceil + 1) \frac{1}{\mu}$. If the next customer arrives within the waiting time, the tagged customer's sojourn time $W(m, 0)$ is $2 \frac{1}{\mu}$. If the next customer does not arrive within the waiting time, the situation under consideration reduces to state $(0, 0)$, and the tagged customer's sojourn time $W(m, 0)$ is $2 \frac{1}{\mu} + \frac{1}{\lambda + \omega}$. Therefore, the expected sojourn time for tagged customers is

$$\begin{aligned} \mathbb{E}[W(m, 0)] &= \Pr(T < T_m)W(m, 0) + \Pr(T \geq T_m)W(m, 0) \\ &= \left(1 - e^{-\lambda \frac{1}{\mu}}\right) \left(2 \frac{1}{\mu}\right) + e^{-\lambda \frac{1}{\mu}} \left(2 \frac{1}{\mu} + \frac{1}{\lambda + \omega}\right). \end{aligned} \quad (9)$$

The customer's expected payoff $U_0(m)$ is

$$\begin{aligned} U_0(m) &= \Pr(T < T_m)(\mathbb{E}[v_0] - C_w W(m, 0)) + \Pr(T \geq T_m)(\mathbb{E}[v_0 - C_r T_w] - C_w W(m, 0)) \\ &= \left(1 - e^{-\lambda \frac{1}{\mu}}\right) \left(V - \alpha P - 2C_w \frac{1}{\mu}\right) \\ &\quad + e^{-\lambda \frac{1}{\mu}} \left(\frac{\lambda}{\lambda + \omega}(V - \alpha P) + \frac{\omega}{\lambda + \omega}(V - P) - C_r \frac{1}{\omega} - C_w \frac{1}{\lambda + \omega} - 2C_w \frac{1}{\mu}\right). \end{aligned} \quad (10)$$

Case 4. The system state is $(m, 0)$ and $m \geq S$. In this case, the customer must wait until service for the first m batches is completed, the waiting time is T_m , where $T_m = (\lceil \frac{m}{S} \rceil + 1) \frac{1}{\mu}$. If the next customer arrives at the system during the waiting time, the tagged customer can immediately match with that customer and be served in the next round. The tagged customer's sojourn time $W(m, 0)$ is $(\lceil \frac{m}{S} \rceil + 1) \frac{1}{\mu}$. If the next customer does not arrive at the system within the waiting time, the state changes to $(0, 0)$, and the customer's sojourn time $W(m, 0)$ is $(\lceil \frac{m}{S} \rceil + 1) \frac{1}{\mu} + \frac{1}{\lambda + \omega}$. Therefore, the expected sojourn time for tagged customers is

$$\begin{aligned} \mathbb{E}[W(m, 0)] &= \Pr(T < T_m)W(m, 0) + \Pr(T \geq T_m)W(m, 0) \\ &= \left(1 - e^{-\lambda \lceil \frac{m}{S} \rceil \frac{1}{\mu}}\right) \left(\left(\lceil \frac{m}{S} \rceil + 1\right) \frac{1}{\mu}\right) + e^{-\lambda \lceil \frac{m}{S} \rceil \frac{1}{\mu}} \left(\left(\lceil \frac{m}{S} \rceil + 1\right) \frac{1}{\mu} + \frac{1}{\lambda + \omega}\right). \end{aligned} \quad (11)$$

The customer's expected payoff $U_0(m)$ is

$$\begin{aligned} U_0(m) &= \Pr(T < T_m)(\mathbb{E}[v_0] - C_w W(m, 0)) + \Pr(T \geq T_m)(\mathbb{E}[v_0 - C_r T_w] - C_w W(m, 0)) \\ &= \left(1 - e^{-\lambda \lceil \frac{m}{S} \rceil \frac{1}{\mu}}\right) \left(V - \alpha P - C_w \left(\lceil \frac{m}{S} \rceil + 1\right) \frac{1}{\mu}\right) \\ &\quad + e^{-\lambda \lceil \frac{m}{S} \rceil \frac{1}{\mu}} \left(\frac{\lambda}{\lambda + \omega}(V - \alpha P) + \frac{\omega}{\lambda + \omega}(V - P) - C_r \frac{1}{\omega} - C_w \frac{1}{\lambda + \omega} - C_w \left(\lceil \frac{m}{S} \rceil + 1\right) \frac{1}{\mu}\right). \end{aligned} \quad (12)$$

Considering the four scenarios above, if a customer arrives while the system is busy, the resulting sojourn time is

$$W(m, j) = \begin{cases} \left\lceil \frac{m+1}{S} \right\rceil \frac{1}{\mu}, & j = 0 \\ \left(\left(\lceil \frac{m}{S} \rceil + 1\right) \frac{1}{\mu}\right) + e^{-\lambda \lceil \frac{m}{S} \rceil \frac{1}{\mu}} \frac{1}{\lambda + \omega}, & j = 1 \end{cases} \quad (13)$$

The expected payoff for the tagged customers is

$$U_0(m) = \left(1 - e^{-\lambda \lceil \frac{m}{S} \rceil \frac{1}{\mu}}\right)(V - \alpha P) + e^{-\lambda \lceil \frac{m}{S} \rceil \frac{1}{\mu}} \left(\frac{\lambda}{\lambda + \omega} (V - \alpha P) + \frac{\omega}{\lambda + \omega} (V - P) - C_r \frac{1}{\omega} - C_w \frac{1}{\lambda + \omega} \right) - C_w \left(\lceil \frac{m}{S} \rceil + 1 \right) \frac{1}{\mu}. \quad (14)$$

$$U_1(m) = V - (1 - \alpha)P - C_w \lceil \frac{m+1}{S} \rceil \frac{1}{\mu}. \quad (15)$$

Therefore, the optimal threshold strategy in state (m, j) is given by Theorem 2.

Theorem 2 (Equilibrium threshold strategy). *When a tagged customer arrives and observes state $(m, 1)$, the customer decides whether to join the server based on the observed number of customers in line. For a given α , the optimal threshold strategy for state $(m, 1)$ is as follows.*

$$m_1^e = S \left\lfloor \frac{(V - (1 - \alpha)P)\mu}{C_w} \right\rfloor - 1. \quad (16)$$

When $m \leq m_1^e$, customers will choose to join the queue. Otherwise, when $m > m_1^e$, customers will balk.

When a tagged customer arrives and observes state $(m, 0)$, before receiving service, he/she becomes an unmatched customer who must wait until the first m batches of customers have completed service. Assuming $U_0(0) > 0$, to ensure that the first customer can join the queue. For a given α , the optimal threshold strategy for state $(m, 0)$ is as follows.

$$m_0^e = \max\{m > 0 : U_0(m) \geq 0\}. \quad (17)$$

When $m \leq m_0^e$, $U_0(m_0^e) \geq 0$, the customer chooses to join the queue. When $m > m_0^e$, $U_0(m_0^e) < 0$, the customer decides not to join the queue.

Proof of Theorem 2. Based on the above discussion, the expected payoff $U_1(m)$ of the tagged customer in state $(m, 1)$ is given by Equation (15), and the key equilibrium condition for customers to choose to join is a non-negative expected payoff. Since $U_1(m)$ is strictly monotonically decreasing with respect to m , there exists a unique integer threshold m_1^e ,

$$m_1^e = \max\{m > 0 : U_1(m) \geq 0\} = S \left\lfloor \frac{(V - (1 - \alpha)P)\mu}{C_w} \right\rfloor - 1. \quad (18)$$

When $m \leq m_1^e$, customers who observe the state $(m, 1)$ will choose to join the queue. When $m > m_1^e$, customers will balk.

By substituting ω^* for ω in Equation (14), we obtain the expected payoff $U_0(m)$ of the tagged customer under state $(m, 0)$, expressed as follows.

$$U_0(m) = \left(1 - e^{-\lambda \lceil \frac{m}{S} \rceil \frac{1}{\mu}}\right)(V - \alpha P) + e^{-\lambda \lceil \frac{m}{S} \rceil \frac{1}{\mu}} \left(\frac{\lambda}{\lambda + \omega^*} (V - \alpha P) + \frac{\omega^*}{\lambda + \omega^*} (V - P) - C_r \frac{1}{\omega^*} - C_w \frac{1}{\lambda + \omega^*} \right) - C_w \left(\lceil \frac{m}{S} \rceil + 1 \right) \frac{1}{\mu}. \quad (19)$$

We prove the existence and uniqueness of the threshold m_0^e in the following three steps.

Step 1. Constructing a continuous function $h(t)$. Let $A = V - \alpha P$, $B = \frac{\omega^*}{\omega^* + \lambda}(V - P) + \frac{\lambda}{\omega^* + \lambda}A - \frac{C_w}{\omega^* + \lambda} - \frac{C_r}{\omega^*}$, and $L = \frac{1}{\mu}$. Then Equation (19) can be rewritten as follows.

$$U_0(m) = \left(1 - e^{-\lambda L \lceil \frac{m}{S} \rceil}\right)A + e^{-\lambda L \lceil \frac{m}{S} \rceil}B - C_w L \left(\lceil \frac{m}{S} \rceil + 1\right). \quad (20)$$

For further analysis, $U_0(m)$ is regarded as a function $h(t)$ of a continuous variable. Let $\frac{t}{S} = \lceil \frac{m}{S} \rceil$, then the expression of $h(t)$ is as follows.

$$h(t) = \left(1 - e^{-\lambda L \frac{t}{S}}\right)A + e^{-\lambda L \frac{t}{S}}B - C_w L \left(\frac{t}{S} + 1\right). \quad (21)$$

From the definition, we have $A > B > 0$ and $L > 0$, because $A - B$ has the following expression,

$$\begin{aligned} A - B &= \frac{\omega^*}{\omega^* + \lambda}A + \frac{C_w}{\omega^* + \lambda} + \frac{C_r}{\omega^*} - \frac{\omega^*}{\omega^* + \lambda}(V - P) \\ &= \frac{\omega^*}{\omega^* + \lambda}(P - \alpha P) + \frac{C_w}{\omega^* + \lambda} + \frac{C_r}{\omega^*} > 0, \end{aligned} \quad (22)$$

so $A > B > 0$ and because $\mu > 0$, $L = \frac{1}{\mu} > 0$. In what follows, we assume that $h(0) = B - C_w L > 0$, which ensures that the first customer arriving at the system is willing to join (i.e., customers are willing to join and wait when the server is idle).

Step 2. Prove that $h(t)$ is strictly quasiconcave and has a unique zero. To determine the monotonicity of $h(t)$, the following two expressions are obtained by solving the first and second derivatives of $h(t)$.

$$h'(t) = \lambda L \frac{1}{S} e^{-\lambda L \frac{t}{S}} (A - B) - C_w \frac{L}{S}. \quad (23)$$

$$h''(t) = -\left(\lambda L \frac{1}{S}\right)^2 e^{-\lambda L \frac{t}{S}} (A - B). \quad (24)$$

Since $A > B$, $h''(t) < 0$ for all $t \geq 0$, so $h'(t)$ is strictly decreasing on $[0, +\infty)$. Furthermore, $\lim_{t \rightarrow +\infty} h'(t) = -C_w \frac{L}{S} < 0$. The sign of $h'(0)$ is discussed in the following two cases.

Case 1. If $h'(0) = \lambda L \frac{1}{S} (A - B) - C_w \frac{L}{S} < 0$, then for all $t \in [0, +\infty)$, $h'(t) < 0$, that is, $h(t)$ is strictly decreasing on $[0, +\infty)$.

Case 2. If $h'(0) = \lambda L \frac{1}{S} (A - B) - C_w \frac{L}{S} \geq 0$, there exists a unique point t_0 such that $h'(t_0) = 0$. When $t > t_0$, $h'(t) < 0$; when $t < t_0$, $h'(t) > 0$, i.e., $h(t)$ is strictly increasing on $[0, t_0]$ and strictly decreasing on $[t_0, +\infty)$.

Since $\lim_{t \rightarrow +\infty} h(t) = -\infty$ in both cases, and $h(0) > 0$ by assumption, according to the intermediate value theorem, $h(t) = 0$ has at least one solution. In case 1, the solution is unique by strict monotonicity. In case 2, because $h'(0) > 0$ and $h(t)$ is strictly increasing on $[0, t_0]$, $h(t)$ remains positive on $[0, t_0]$, and the unique zero solution is in the decreasing interval $(t_0, +\infty)$. Thus, in both cases, there exists a unique point $t^e > 0$ such that $h(t^e) = 0$. When $t < t^e$, $h(t) > 0$, and when $t > t^e$, $h(t) < 0$.

Step 3. Prove the uniqueness and existence of discrete thresholds. After obtaining the unique continuous threshold t^e , we return to the discrete setting. Although $U_0(m)$ is ladder-shaped due to the existence of the upper integer operation, its continuous function $h(t)$ is strictly quasiconcave and has a unique zero point t^e . From the established continuous mapping relation $U_0(m) = h(S \lceil \frac{m}{S} \rceil)$, the maximum integer m satisfying $U_0(m) \geq 0$ can be directly inverse solved by $S \lceil \frac{m}{S} \rceil \leq t^e$, set $m_0^e = S \lfloor \frac{t^e}{S} \rfloor$. Thus $U_0(m_0^e) \geq 0$ and $U_0(m_0^e + 1) < 0$. Therefore, m_0^e is the threshold strategy for any customer who observes the state $(m, 0)$ upon arrival. In case 2 of step 2 discussed earlier, if there are already more customers in the queue but the queue length does not exceed m_0^e , customers observing state

$(m, 0)$ have a stronger incentive to join the queue, as the benefit function increases for a period of time before decreasing. This feature differs from similar models in the Naor [1] research school, where the payoff function strictly decreases with queue length, and the threshold strategy is more pronounced. This completes the proof. $\square$

Remark 1. Combining the preceding results, the equilibrium strategy adopted by customers is characterized by the group (m_0^e, m_1^e) . A customer who observes system state (m, j) joins the queue if $m < m_1^e$ and balks otherwise. Since $m_0^e < m_1^e$, the maximum queue length is $2m_0^e$.

Remark 2. Unlike the M/M/1 batch service queueing model studied by Bountali [11] and Nguyen [12], existing literature assumes equal cost sharing among customers in the same batch. However, this assumption ignores the asymmetry in waiting time between the batch initiator and subsequent joiners. Therefore, this work introduces a proportional pricing coefficient, α , to make the cost-sharing mechanism more realistic. On the other hand, compared with M/M/1 systems, the M/M/S architecture is more common and necessary in practice. The M/M/1 system is vulnerable, and once the server fails, the overall system will crash and become completely blocked. In view of this, the work introduces S access points (multiple access points) to improve the model's applicability to real-world operational settings and the system's robustness. The effectiveness of the above improvements is verified in the numerical experiments in Section 5, and the results in Figure 5 clearly demonstrate the significant performance advantages of the current model over Nguyen [12] and the fixed-batch M/M/3 model.

3.3. Calculation of Thresholds

In the observable batch-service queueing model, the equilibrium strategy of customers is fully characterized by two thresholds, m_0^e and m_1^e . The threshold m_1^e is given by Equation (16). Since Equation (17) corresponding to the threshold m_0^e cannot be analytically solved, this work employs a binary searching algorithm to solve it on the continuous domain, and then rounds it to obtain the integer threshold.

The continuous expected payoff function when the customer observes the state $(m, 0)$ is given by Equation (21). Let the lower bound of the search interval be $L_b = 1$, and the upper bound be initialized to $U_b = 2$. If $h(U_b) > 0$, it indicates that U_b is not the upper bound of the threshold. Set $U_b = 2U_b$ and repeat the judgment until $h(U_b) \leq 0$. From the proof of Theorem 2, we can see that $h(t)$ is strictly quasiconcave and $\lim_{t \rightarrow +\infty} h(t) = -\infty$, so there must exist a unique $t^e \in [L_b, U_b]$ such that $h(t^e) = 0$. Based on this, the standard binary search is performed on the interval $[L_b, U_b]$, and the continuous zero t^e is approximated to the accuracy of $1e - 6$. Finally, the discrete threshold is obtained by $m_0^e = \lfloor \frac{t^e}{5} \rfloor$. If $h(1) < 0$, then $t^e \in (0, 1)$, and $m_0^e = 0$ is returned directly. Algorithm 1 provides a method for calculating the threshold in state $(m, 0)$.

Algorithm 1 Binary Searching Algorithm for Computing the Threshold m_0^e **Require:** $V, \lambda, \mu, \alpha, P, C_w, C_r, S > 0$ **Ensure:** m_0^e

```

1: if  $h(1) < 0$  then
2:   return 0
3: end if
4:  $L_b \leftarrow 1$ 
5:  $U_b \leftarrow 2$ 
6: while  $h(U_b) > 0$  do
7:    $U_b \leftarrow 2 \times U_b$ 
8: end while
9: while  $U_b - L_b > 1e - 6$  do
10:   $mid \leftarrow (L_b + U_b)/2$ 
11:  if  $h(mid) > 0$  then
12:     $L_b \leftarrow mid$ 
13:  else
14:     $U_b \leftarrow mid$ 
15:  end if
16: end while
17:  $t^e \leftarrow (L_b + U_b)/2$ 
18:  $m_0^e \leftarrow S \lfloor \frac{t^e}{S} \rfloor$ 
19: return  $m_0^e$ 

```

4. Performance Measures in Equilibrium*4.1. System Modeling and Steady-State Distribution Under the Equilibrium State*

In the equilibrium state, customers follow the threshold strategy (m_0^e, m_1^e) to make decisions. To characterize the system dynamics, a continuous time Markov chain $\{(X(t), Y(t)) : t \geq 0\}$ is established, where $X(t)$ represents the number of customers waiting in the queue at time t , and $Y(t)$ represents the service mode at time t . The state space is defined as: $\mathcal{S} = \{(x, y) | x \in \{0, 1, \dots, N\}, y \in \mathcal{Y}_S\}$, where $N = 2m_0^e$ and $\mathcal{Y}_S = \{0, 1\} \cup \{2, 4, 6, \dots, 2S\}, S > 1$.

Let $p_{x,y}$ denote the steady-state probability of the state $(x, y) \in \mathcal{S}$. Clearly, for $p_{x,y} = 0$ when $x > 2m_0^e$, the balance equations for the steady-state probability distribution are given as follows.

When $y = 0$:

$$\begin{aligned} \lambda p_{0,0} &= \mu p_{0,1} + \mu \left(\sum_{i=1}^S p_{0,2i} \right), \\ (\lambda + \omega^*) p_{1,0} &= \lambda p_{0,0} + \mu p_{1,1} + \mu \left(\sum_{i=1}^S p_{1,2i} \right). \end{aligned} \quad (25)$$

When $y = 1$:

$$\begin{aligned} (\lambda + \mu) p_{0,1} &= \omega^* p_{1,0}, \\ (\lambda + \mu) p_{x,1} &= \lambda p_{x-1,1}, 1 \leq x \leq N-1, \\ \mu p_{N,1} &= \lambda p_{N-1,1}. \end{aligned} \quad (26)$$

When $y = 2$:

$$\begin{aligned}(\lambda + \mu)p_{0,2} &= \lambda p_{1,0} + \mu p_{2,1} + \mu \left(\sum_{i=1}^S p_{2,2i} \right), \\(\lambda + \mu)p_{1,2} &= \lambda p_{0,2} + \mu p_{3,1} + \mu \left(\sum_{i=1}^S p_{3,2i} \right), \\(\lambda + \mu)p_{x,2} &= \lambda p_{x-1,2}, 2 \leq x \leq N-1, \\ \mu p_{N,2} &= \lambda p_{N-1,2}.\end{aligned}\tag{27}$$

When $y = 4, 6, \dots, 2S-2$:

$$\begin{aligned}(\lambda + \mu)p_{0,y} &= \mu p_{y,1} + \mu \left(\sum_{i=1}^S p_{y,2i} \right), \\(\lambda + \mu)p_{1,y} &= \lambda p_{0,y} + \mu p_{y+1,1} + \mu \left(\sum_{i=1}^S p_{y+1,2i} \right), \\(\lambda + \mu)p_{x,y} &= \lambda p_{x-1,y}, 2 \leq x \leq N-1, \\ \mu p_{N,y} &= \lambda p_{N-1,y}.\end{aligned}\tag{28}$$

When $y = 2S$:

$$\begin{aligned}(\lambda + \mu)p_{0,2S} &= \mu p_{2S,1} + \mu \left(\sum_{i=1}^S p_{2S,2i} \right), \\(\lambda + \mu)p_{x,2S} &= \lambda p_{x-1,2S} + \mu p_{x+2S,1} + \mu \left(\sum_{i=1}^S p_{x+2S,2i} \right), 1 \leq x \leq N-2S, \\(\lambda + \mu)p_{x,2S} &= \lambda p_{x-1,2S}, N-2S+1 \leq x \leq N-1, \\ \mu p_{N,2S} &= \lambda p_{N-1,2S}.\end{aligned}\tag{29}$$

The linear system, together with the normalization condition, uniquely determines the steady-state probability distribution, laying the foundation for subsequent performance-measure calculations and equilibrium analysis.

$$p_{0,0} + p_{1,0} + \sum_x p_{x,1} + \sum_{i=1}^S \left(\sum_x p_{x,2i} \right) = 1.\tag{30}$$

Let $\mathcal{Y} = \{1, 2, 4, 6, \dots, 2S\}$ denote the set of possible numbers of customers in service, and let $M = |\mathcal{Y}| = S + 1$. For each $y \in \mathcal{Y}$, define the probability vector $\mathbf{p}_y$ as follows.

$$\mathbf{p}_y = [p_{0,y}, p_{1,y}, \dots, p_{N,y}] \in \mathbb{R}^{N+1}.\tag{31}$$

The boundary probability $p_{0,0}$ is treated separately. The global probability vector is

$$\mathbf{p} = [p_{0,0}, p_{1,0}, \mathbf{p}_1^T, \mathbf{p}_2^T, \mathbf{p}_4^T, \dots, \mathbf{p}_{2S}^T]^T \in \mathbb{R}^{(N+1)M+2}.\tag{32}$$

Then, the balance equation can be expressed in Block Tridiagonal form as follows.

$$\mathbf{H}\mathbf{p} = \mathbf{b},\tag{33}$$

where $\mathbf{H} \in \mathbb{R}^{((N+1)M+2) \times ((N+1)M+2)}$ is the coefficient matrix, and $\mathbf{b} \in \mathbb{R}^{(N+1)M+2}$ is the right-hand-side vector. The matrix $\mathbf{H}$ has the following block structure,

$$\mathbf{H} = \begin{pmatrix} B_{0,0} & B_{0,1} & B_{0,2} & B_{0,4} & \cdots & B_{0,2S} \\ B_{1,0} & B_{1,1} & B_{1,2} & 0 & \cdots & 0 \\ B_{2,0} & B_{2,1} & B_{2,2} & B_{2,4} & \cdots & 0 \\ 0 & B_{4,1} & B_{4,2} & B_{4,4} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & B_{2S,2S} \end{pmatrix}, \quad (34)$$

where each $B_{i,j}$ is a submatrix derived from the transition rates between service states i and j . The dimensions of these submatrices depend on the number of queue-length states associated with each service state, as specified below.

- $B_{0,0} \in \mathbb{R}^{2 \times 2}$, corresponding to the two boundary states $(0,0)$ and $(1,0)$.
- $B_{0,y} \in \mathbb{R}^{2 \times (N+1)}$ for $y \neq 0$, corresponding to the coupling between the boundary layer and the general service states.
- $B_{x,0} \in \mathbb{R}^{(N+1) \times 2}$ for $x \neq 0$, corresponding to the coupling from the general service states back to the boundary layer.
- All other non-zero blocks are in $\mathbb{R}^{(N+1) \times (N+1)}$, corresponding to the transitions among the general service states.

Specifically, the diagonal blocks $B_{y,y}$ contain the diagonal terms $\lambda + \mu$ for $y \in \mathcal{Y}$, with the following exceptions.

- The block associated with $p_{0,0}$ uses λ instead of $\lambda + \mu$.
- The block associated with $p_{1,0}$ uses $\lambda + \omega^*$ instead of $\lambda + \mu$.
- The blocks associated with $p_{N,y}$ for $y \in \mathcal{Y}$ use μ instead of $\lambda + \mu$.

These diagonal blocks also contain the off-diagonal terms arising from the balance equations between adjacent x -states. The off-diagonal blocks $B_{i,j}$ with $i \neq j$ capture the coupling effects induced by service completions that alter the number of busy servers. The first block row $B_{0,y}$ encodes the boundary Equation (25) for $p_{0,0}$.

The right-hand side vector $\mathbf{b}$ has only one non-zero entry: the first component equals $\lambda p_{0,0}$, which appears in the boundary balance equation for $p_{1,0}$. All other components are zero. The normalization condition $p_{0,0} + p_{1,0} + \sum_{y \in \mathcal{Y}} \sum_x p_{x,y} = 1$ is imposed after each iteration to ensure that the computed probabilities form a valid stationary distribution.

The coefficient matrix $\mathbf{H}$ is nonsingular due to the irreducibility and positive recurrence of the underlying CTMC. It is a large, sparse matrix, and the coupling between blocks prevents a closed-form solution. We therefore solve the system numerically using the Gauss–Seidel iteration described in Algorithm 2. Its convergence is guaranteed by the M-matrix property of the CTMC generator.

To illustrate the system dynamics described by the equilibrium equations, Figure 2 shows a typical example of a state transition diagram, where $S = 3$. The diagram is directly generated by the above general transition rules, and the general state space and equilibrium equation are applicable to any S .

Algorithm 2 Steady-State Probability Solution Based on the Gauss–Seidel Iteration

Require: $\lambda, \mu, \omega^*, M, N, \varepsilon, K_{\max} > 0, \theta \in (1, 2)$
Ensure: $\mathbf{p} = [p_{0,0}, p_{1,0}, \mathbf{p}_1^T, \mathbf{p}_2^T, \mathbf{p}_4^T, \dots, \mathbf{p}_{2S}^T]^T$

- 1: Construct $\mathbf{H}$ according to Equations (25) ~ (29) and (34)
- 2: Construct $\mathbf{b}$ according to Equations (25) ~ (29)
- 3: $D \leftarrow (N + 1)M + 2$
- 4: $\mathbf{p}^{(0)} \leftarrow (1/D) \cdot \mathbf{1}$
- 5: **for** $k \leftarrow 1$ to $K_{\max}$ **do**
- 6: $\mathbf{p}_{\text{old}} \leftarrow \mathbf{p}^{(k-1)}$
- 7: **for** $i \leftarrow 1$ to D **do**
- 8: $\text{sum} \leftarrow \mathbf{b}_i - \sum_{j=1}^{i-1} \mathbf{H}_{i,j} \cdot \mathbf{p}_j^{(k)} - \sum_{j=i+1}^D \mathbf{H}_{i,j} \cdot \mathbf{p}_j^{(k-1)}$
- 9: $\mathbf{p}_i^{(k)} \leftarrow \text{sum} / \mathbf{H}_{i,i}$
- 10: **end for**
- 11: $\mathbf{p}^{(k)} \leftarrow \theta \mathbf{p}^{(k)} + (1 - \theta) \mathbf{p}_{\text{old}}$
- 12: $\mathbf{p}^{(k)} \leftarrow \mathbf{p}^{(k)} / \sum_i \mathbf{p}_i^{(k)}$
- 13: $r^{(k)} \leftarrow \|\mathbf{H} \mathbf{p}^{(k)} - \mathbf{b}\|_{\infty}$
- 14: **if** $r^{(k)} < \varepsilon$ **then**
- 15: **return** $\mathbf{p}^{(k)}$
- 16: **end if**
- 17: **end for**
- 18: **return** $\mathbf{p}^{(K_{\max})}$ with warning

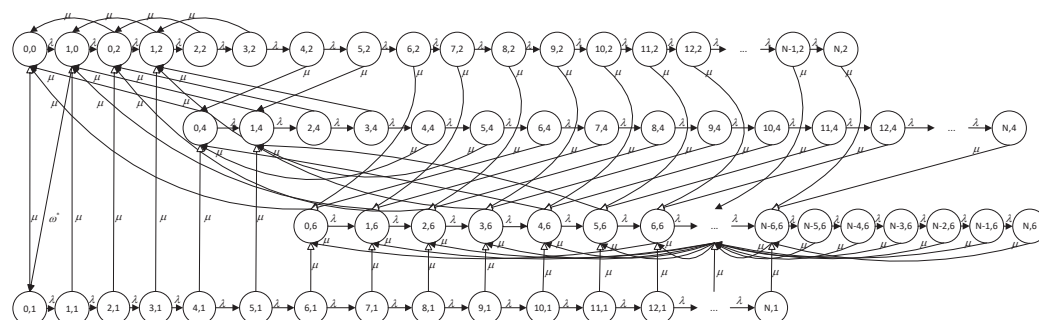


Figure 2. Transition rate graph in equilibrium ($S = 3$).

Remark 3. The system state (m, j) is the logical perspective from which customers make decisions, while the Markov state (x, y) is the physical characterization of the system's operation. The physical CTMC state (x, y) is mapped to the strategic state (m, j) by

$$m = \left\lceil \frac{x + y}{2} \right\rceil, \quad j = x \bmod 2. \quad (35)$$

A refined mapping that explicitly tracks the internal batch structure is provided in Appendix A. As shown there, the equilibrium threshold strategy derived from the coarse mapping $(x, y) \rightarrow (m, j)$ is identical to that obtained from the refined mapping. Therefore, the use of (m, j) in the main text involves no loss of generality.

4.2. Quantification of Performance Measures

This section quantifies the model's key performance measures, derives their mathematical expressions, and lays the foundation for subsequent numerical experiments.

Customers arriving at the system decide whether to join according to the equilibrium threshold strategy established in Theorem 2. A tagged customer who observes state $(m, 0)$

joins if and only if $m \leq m_0^e$; a tagged customer who observes state $(m, 1)$ joins if and only if $m \leq m_0^e$. Therefore, the complete set of balking states is as follows.

$$\mathcal{B} = \{(x, y) : j = 0, m > m_0^e\} \cup \{(x, y) : j = 1, m > m_1^e\}. \quad (36)$$

Since $j = x \bmod 2$ and $m = \lceil \frac{x+y}{2} \rceil$, the complete set of balking states can be rewritten as follows.

$$\mathcal{B} = \left\{ (x, y) : j = 0, \lceil \frac{x+y}{2} \rceil > m_0^e \right\} \cup \left\{ (x, y) : j = 1, \lceil \frac{x+y}{2} \rceil > m_1^e \right\}. \quad (37)$$

The blocking probability is the probability that an arriving customer is refused entry to the system. Therefore, the blocking probability can be expressed as

$$P_b = \sum_{(x,y) \in \mathcal{B}} p_{x,y}. \quad (38)$$

Since $m_0^e < m_1^e$, the capacity constraint is governed by m_0^e : The system accommodates at most m_0^e complete batches. The CTMC state space is thus truncated to $x \leq 2m_0^e$. Every balking state satisfies $x > 2m_0^e$, and hence lies outside the truncated state space. In numerical computation, the blocking probability is equivalently obtained from the boundary probabilities,

$$P_b = \sum_y p_{2m_0^e, y}. \quad (39)$$

Equation (39) is used in numerical experiments; the equivalence between Equations (38) and (39) follows from the truncation $x \leq 2m_0^e$, under which all balking states are mapped to the boundary. Equivalently, let $\lambda_e = \lambda(1 - P_b)$ denote the effective arrival rate, i.e., the actual rate at which customers successfully enter the system. Then the blocking probability can also be expressed as $P_b = 1 - \lambda_e/\lambda$, which is the standard queueing-theoretic characterization. This equivalent form will be used in the resource utilization equation below. The two expressions are consistent because, in the truncated CTMC, the total flow that successfully enters the system is exactly $\lambda(1 - \sum_y p_{2m_0^e, y})$.

When the customer arrives, the observed system state is $(m, 1)$, meaning there are m complete batches in the system and an odd number of customers in the current queue. At this time, the individual expected payoff $U_1(m)$ of state $(m, 1)$ is given by Equation (15).

When the customer arrives, the observed system state is $(m, 0)$, indicating that there are m complete batches in the system and no unmatched customers in the current queue. The customer becomes the unmatched customer and must wait for the next customer to arrive to form a batch, or use the service alone. Therefore, the individual expected payoff $U_0(m)$ of state $(m, 0)$ is given by Equation (19).

Resource utilization rate ρ is defined as the ratio of the effective arrival rate to the maximum service capacity. Therefore, the resource utilization equation is as follows.

$$\rho = \frac{\lambda_e}{S \cdot \mu \cdot 2} = \frac{\lambda(1 - P_b)}{2S\mu}. \quad (40)$$

The total revenue generated by the system per unit time is given by

$$I_a = \lambda_e V = (1 - P_b) \lambda V. \quad (41)$$

Let Q be the average number of customers in the system,

$$Q = \sum_x \sum_y (x + y) p_{x,y}. \quad (42)$$

The total social cost consists of three components: the customer waiting cost $C_w \cdot Q$, the service operating cost $C_o \cdot \mu\rho$, where C_o is the operating cost per completed batch service, and the capacity cost $C_s \cdot S$, where C_s is the fixed cost per access point per unit time. Therefore, the social welfare, defined as the total service value minus all social costs, is given by

$$SW = (1 - P_b)\lambda V - C_w Q - C_o \mu\rho - C_s S. \quad (43)$$

4.3. Optimization of Pricing Parameters Based on Social Welfare

The social welfare function $SW(\alpha)$ is nonconvex and nonlinear, making traditional gradient methods difficult to apply. The essence of this problem is a Stackelberg game, and the following two-layer optimization model is constructed: The upper-level social planner chooses α for the observed exogenous demand level λ to maximize SW ; the lower-level customers engage in noncooperative games under given parameters to achieve the Nash equilibrium threshold strategy. This work uses a grid search to find the optimal proportional pricing coefficient α in the one-dimensional parameter space, given an exogenous arrival rate λ . This approach effectively handles the non-smooth operations in the lower-level solution and accurately identifies the global optimum.

$$\begin{aligned} \max_{\alpha} SW(\alpha) &= (1 - P_b)\lambda V - C_w Q - C_o \mu\rho - C_s S \\ \text{s.t.} \quad &\begin{cases} \omega^* = \arg \max_{\omega \geq 0} f(\omega) \\ m_1^e = \max \left\{ m > 0 : V - (1 - \alpha)P - C_w \left\lceil \frac{m+1}{S} \right\rceil \frac{1}{\mu} \geq 0 \right\} \\ m_0^e = \max \{ m > 0 : U_0(m) \geq 0 \} \\ 0 < \alpha \leq 0.5, \lambda, \mu, C_r, C_w, S, C_o, C_s > 0, V > P > 0 \end{cases} \end{aligned} \quad (44)$$

where $f(\omega)$ is given by Equation (2), and $U_0(m)$ is given by Equation (19).

5. Numerical Experiments

To verify the theoretical validity of the strategic batch queuing model and evaluate the overall performance of differential pricing and a multi-server architecture, we conducted a series of numerical experiments. The common parameters are set as follows: $\mu = 4$, $V = 8$, $P = 5$, $C_w = 4$, $C_r = 2$, $C_o = 1$, $C_s = 1$, $\lambda = 20$, $S = 3$ for the current model and M/M/3 batch model, and $S = 1$ for the Nguyen model [12]. The experiment relies on the grid search algorithm to solve the nonconvex and nonsmooth social welfare optimization problem, compares the social welfare and resource utilization between the current model and the two benchmark models, and reveals the sensitivity of the optimal proportional pricing coefficient α . All conclusions are of universal significance; under various parameter configurations, repeated experiments can yield consistent results.

5.1. Model Parameter Optimization and Parameter Sensitivity Analysis

To examine the single-factor effect of α on social welfare and determine its optimal value, grid search is used to scan within $\alpha \in [0.01, 0.5]$, fixing $\mu = 4$, $V = 8$, $P = 5$, $C_w = 4$, $C_r = 2$, $C_o = 1$, $C_s = 1$, $S = 3$, and drawing the curve of social welfare changing with α . As shown in Figure 3, the overall social welfare shows a stepwise increase as α increases, but the growth rate gradually slows down. When α is in the low range of 0 to 0.15, SW remains at 96.5; When α reaches 0.41 or above, SW stabilizes at 103, an increase of about 6.7% from the lowest value. Further increasing α thereafter no longer improves in social welfare, and the global optimal $\alpha = 0.41$.

Proportional pricing coefficient α , as the key control tool of the platform, has a two-way effect on customer participation costs and platform revenue: when α is too small, the platform does not get enough revenue from unit services and social welfare is low; When α exceeds 0.41, the marginal contribution of further price increases to platform revenue is offset by customer churn, and social welfare enters the saturation range. To achieve continuous optimization of social welfare, we suggest that the platform abandon the traditional average-pricing model and adopt a time-dependent dynamic pricing strategy. α will be moderately lowered during periods of low customer flow, and the threshold of customer participation will be lowered to stimulate customer flow. During periods of high customer flow, α will be slightly adjusted to balance the supply and demand relationship of the system.

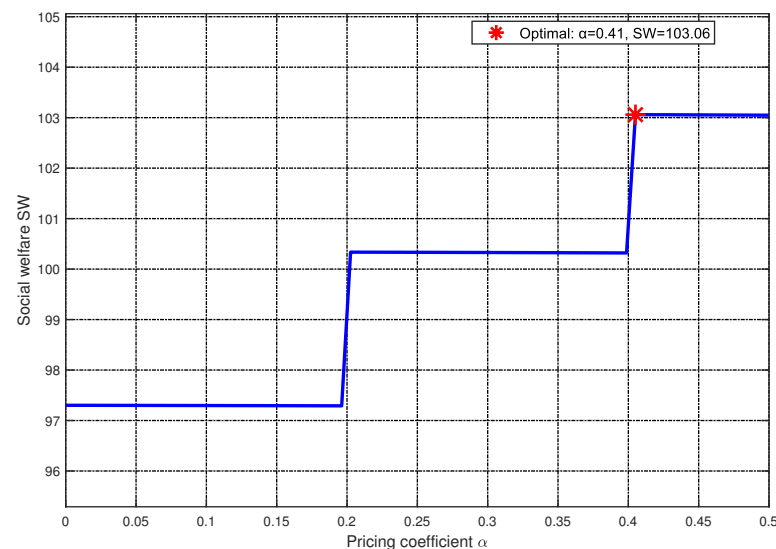


Figure 3. Curve of Social Welfare with α .

To examine the robustness of the platform's optimal proportional pricing coefficient α^* , the reservation fee C_r , the number of access points S , the fixed operating cost C_o and the unit-time cost per access point C_s are changed separately, and the sensitivity curves of α^* with single parameter are plotted. It can be seen from Figure 4 that when C_r , C_o and C_s change significantly in the range of 0.5 ~ 4, α^* remains stable at 0.41, and the curve is a horizontal straight line. Only the number of access points S has periodic differences. When S is between 1 and 2, α^* rises rapidly with the expansion of access points. When $S \geq 3$, α^* quickly converges to 0.41 and remains constant.

The reservation fee, fixed operating cost, and access point costs will only alter the platform's overall cost structure but will not change the optimal proportional pricing coefficient under supply–demand matching. Only when the supply capacity provided by access points is insufficient will the system's supply and demand constraints compress the reasonable pricing range. Therefore, we recommend that the platform's daily operation does not need to adjust pricing according to the regular fluctuations of reservation fees, access operation and maintenance, and fixed costs. During the platform planning stage, it is advisable to configure at least three access points to reduce price fluctuations caused by supply shortages.

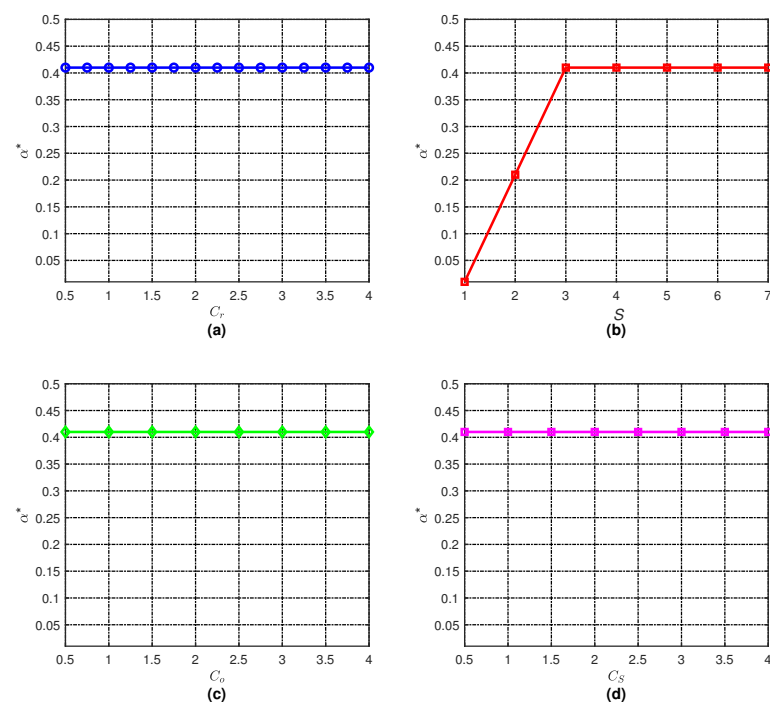


Figure 4. Sensitivity analysis of α^* under different parameters: (a) α^* versus C_r ; (b) α^* versus S ; (c) α^* versus C_o ; (d) α^* versus C_S .

5.2. Performance Comparison of Multiple Models and Analysis of System Load Characteristics

To evaluate the performance advantages of the proposed strategic batch queuing mechanism relative to the traditional model, this experiment uses two classical batch queuing models as benchmarks. The current model sets the optimal proportional pricing coefficient $\alpha = 0.41$ and the number of access points $S = 3$, the M/M/3 batch model sets $\alpha = 0.5$ and $S = 3$, and the Nguyen [12] single server model sets $\alpha = 0.5$ and $S = 1$. The comparison curves of the social welfare for the three models as a function of the customer arrival rate λ are shown. As shown in Figure 5, the welfare peak of the current model reaches 108.9, corresponding to the optimal arrival rate $\lambda = 27$; the peak of the M/M/3 batch model is 107.6, corresponding to $\lambda = 19$, and the welfare decays rapidly after the peak. The peak of the Nguyen model is only 28.6, corresponding to $\lambda = 8$, and the performance is the weakest. The current model carries 42% more customer flow than the fixed-batch model with the same number of access points, and the post-peak attenuation is gentler. Compared with the single-server model, the capacity advantage of multiple access points increases social welfare by nearly 3.8 times.

There is a significant gap in the performance of the three types of models. On the one hand, the strategic waiting mechanism and differentiated pricing strategy introduced in this work can improve welfare, throughput capacity, and customer flow stability with the same number of access points. On the other hand, the capacity advantage of the multi-access points architecture greatly strengthens the system's anti-congestion capability. Therefore, we suggest that, in practice, priority should be given to adding access points, abandoning the traditional forced-batch service mode, and aligning the layout of multiple access points with flexible service rules and an optimized pricing scheme, thereby maximizing social welfare and ensuring the system's long-term stable operation.

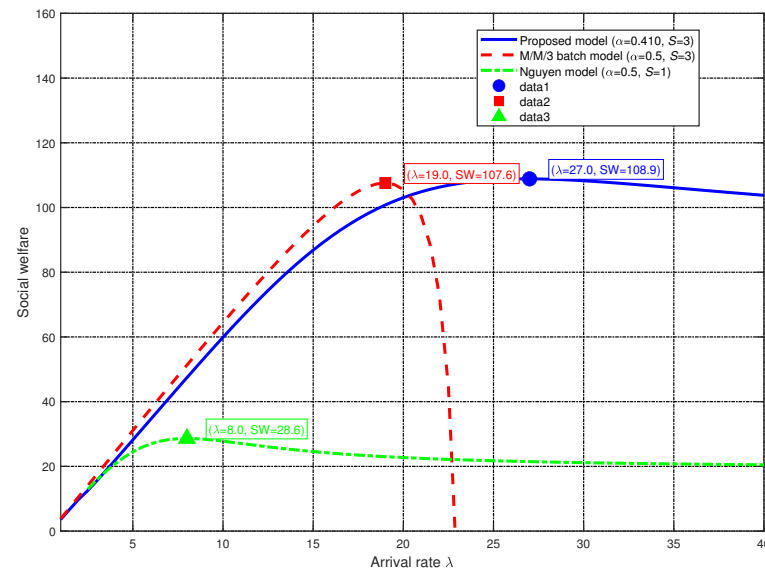


Figure 5. Comparison curves of social welfare performance of three types of queuing models.

The fourth experiment explores the impact of the customer arrival rate on the load status of different queuing systems and verifies the actual control effect of strategic batch processing and a differentiated pricing strategy. The public parameter settings are consistent. The current model sets the optimal proportional pricing coefficient $\alpha = 0.41$ and the number of access points $S = 3$; the M/M/3 batch model sets $S = 3$ and $\alpha = 0.5$; and the Nguyen single-server model sets $S = 1$ and $\alpha = 0.5$. The comparison curves of the resource utilization rates for the three model types as a function of the arrival rate are shown. As shown in Figure 6, the utilization rate of the Nguyen [12] single-server model increases rapidly with the arrival rate, and the system essentially reaches full load at $\lambda \geq 15$. The M/M/3 batch model with the same three-server architecture exhibits a faster load growth rate, and at $\lambda = 25$ the utilization rate reaches 0.99, approaching saturation. The model's utilization rate in this work increases more smoothly and can maintain a service margin of 5~10% in the middle- and high-customer-flow ranges. The utilization rate at $\lambda = 30$ is 0.95, and it is not fully loaded until $\lambda = 40$. The overall load performance of the model is better than that of the other two comparative models.

The differences in performance among the three types of models stem from differences in architecture and control mode. The capacity of a single server is limited, which is prone to congestion. The traditional forced batch processing model lacks the ability to control dynamically. This work integrates the layout of multiple access points, customer strategic waiting, and an optimal pricing strategy to reasonably divert customer flow and delay system saturation. In actual operation, the multi-access point architecture should be preferred; the conventional utilization rate of the system should be controlled within 0.85~0.90%, with 0.95 as the load warning value; and pricing should be adjusted in time. When the arrival rate is $\lambda \geq 30$, we can ensure the system's smooth operation by adding a temporary access point and guiding customers to stagger peak times.

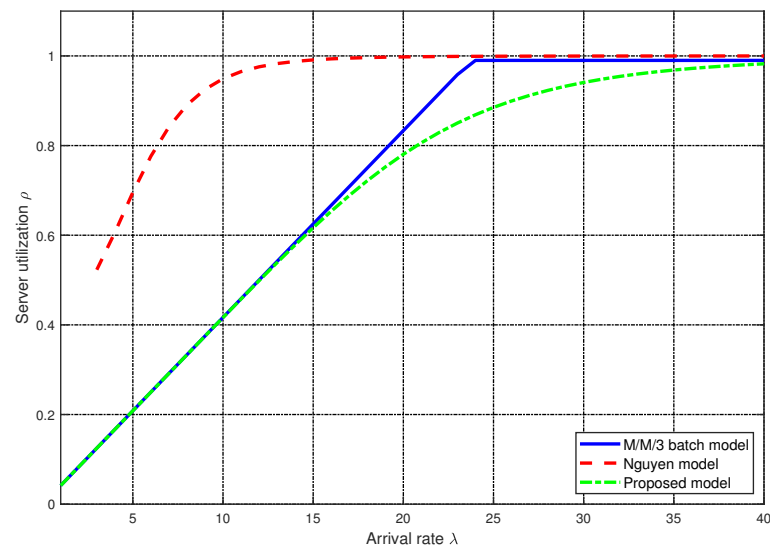


Figure 6. Comparison of resource utilization and load characteristics among three types of queuing models.

5.3. Analysis of the Impact of Batch Size and Number of Access Points on Individual Payoff

The fifth experiment examines the law of change in personal income under the joint effect of batch size and the number of access points S in the current model. The public parameter settings are consistent, and other parameters are set as the optimal parameter sets $\alpha = 0.41$. A comparative line chart containing three curves, $S = 3$, $S = 5$, and $S = 7$ and the zero payoff reference line is drawn. As shown in Figure 7, the relationship between individual payoff and the number of batches depends on the system state: in state $(m, 0)$, as the number of batches in the system increases, the individual payoff first rises slightly and then decreases in a stepwise manner, exhibiting a pattern of constant, segmented decline; in state $(m, 1)$, the individual payoff decreases in a stepwise manner as the number of batches in the system. Compared with the two types of states, $(m, 0)$ has a higher overall payoff, a slower rate of decline, and a larger range of positive payoff. At the same time, under the same system state, the larger the number of access points, the higher the overall position of the payoff curve, and the amplitude of payoff attenuation is relatively smaller. Both functions have a unique threshold point at which the payoff turns negative within their respective domains, and this threshold point shifts to the right as S increases. This is consistent with the theoretical conclusion in Theorem 2 regarding the existence of a unique threshold, and provides numerical verification for Theorem 2.

The continuous backlog in customer flow will increase the burden on system operations, further reducing individual payoff. Once the queue length exceeds the threshold, the waiting costs offset the economies of scale gained from batch processing, resulting in a decline in overall benefits. For different types of arrival states, the rate at which payoff declines and the thresholds vary; the system must take multiple scenarios into account simultaneously and adopt the most stringent of these as the basis for actual control measures. Based on the analysis results, it is necessary to control the number of waiting customers during operations, set the customer flow threshold based on the number of access points, and optimize access point configuration to improve overall payoff performance.

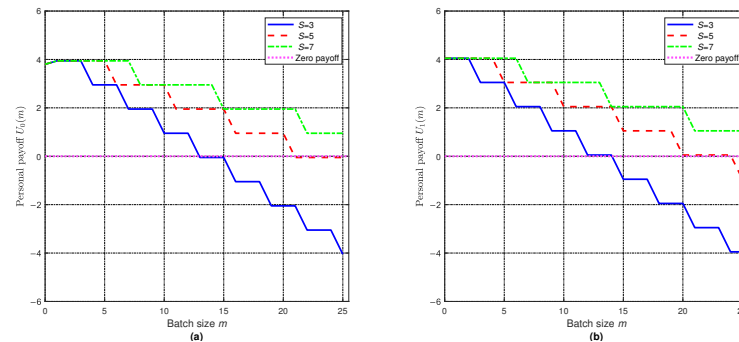


Figure 7. Variation curves of customer payoff with batch size and number of access points under two types of system states: (a) $U_0(m)$ versus m for $S = 3, 5, 7$; (b) $U_1(m)$ versus m for $S = 3, 5, 7$.

6. Conclusions

In this work, we investigate the equilibrium strategy of customers in a multi-server batch service queuing system with a proportional pricing coefficient. We also analyze the impact of the proportional pricing coefficient and the number of access points on customers' expected payoff and social welfare. When the system is available, the optimal waiting strategy is the balance point between consumption and benefit. The expected payoff for customers is negatively correlated with waiting costs and the proportional pricing coefficient, which means that lower waiting costs and a lower proportional pricing coefficient will motivate customers to join the system. When the system is occupied, we derive the expected payoff and equilibrium joining strategy in two different system states. When the queue length exceeds the threshold, customers no longer join, and customers' expected payoff is positively correlated with S . This means that the more access points there are in the system, the easier it is for customers to join. We design a binary searching algorithm to solve the threshold in the $U_0(m)$ state that is difficult to parse and to solve. Finally, we use a grid search algorithm to find the maximum social welfare.

Through extensive numerical experiments, we find that social welfare increases in a stepwise manner as α increases and that there exists an optimal social welfare and an optimal proportional pricing coefficient. Robustness analysis indicates that S is the key variable influencing the optimal proportional pricing coefficient, while the effects of C_0 , C_r , and C_s are negligible. Compared with the M/M/3 fixed-batch model and the Nguyen [12] single-server model, our model achieves the highest social welfare peak and the smoothest upward trend in utilization rate. In addition, increasing the number of access points can significantly improve personal payoff.

However, this model assumes the server is absolutely reliable and does not account for its breakdown and repair mechanisms. In actual batch service systems, equipment breakdowns are common. When a server breaks, it not only disrupts service to customers but also exacerbates queue congestion, significantly altering customer joining decisions and the system's equilibrium state. Future research can introduce unreliable servers to explore the impact of failure and repair rates on customer equilibrium strategies, dynamic pricing, and social welfare.

Author Contributions: Conceptualization, P.W., Y.C. and Z.W.; methodology, P.W., W.P. and Y.C.; software, Y.Y. and Y.C.; validation, P.W. and W.P.; formal analysis, P.W.; writing—original draft preparation, P.W.; writing—review and editing, W.P. and Z.W.; visualization, Y.Y. and Y.C.; supervision, Z.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Guangxi Natural Science Foundation (2026GXNSFAA00640927), the National College Student Innovation and Entrepreneurship Training Program at Guilin University of Electronic Technology (202510595018), the Science and Technology Project of Guangxi (Guike

AD25069086), Innovation Project of Guangxi Graduate Education (JGY2025159, YCSW2025378), Innovation Project of GUET Graduate Education (2024YCXS113), Course Construction Project of GUET Graduate Education (YKC202412).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: The author would like to express sincere gratitude to H.L. for the valuable assistance in the writing of this manuscript.

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

Appendix A. Refined Strategic-State Mapping

The strategic state (m, j) used in the main text is obtained by aggregating the physical CTMC state (x, y) . This appendix introduces a refined quadruple mapping that preserves all information contained in (x, y) and proves that it does not alter the threshold strategy results derived in the main text.

Appendix A.1. Refined Mapping

We define the mapping from the physical state (x, y) to the quadruple (m, j, r, q) as follows.

$$m = \left\lceil \frac{x+y}{2} \right\rceil, \quad j = x \bmod 2, \quad r = \left\lfloor \frac{x}{2} \right\rfloor, \quad q = y \bmod 2. \quad (\text{A1})$$

Here m denotes the total number of batches in the system, where every two customers form one complete batch and any remaining single customer constitutes a separate batch; j denotes the parity of the queue length; r denotes the number of complete batches already formed in the queue; and q indicates whether there is a single customer in service.

Proposition A1. *The mapping $\Phi(x, y) = (m, j, r, q)$ is injective on the set of feasible physical states.*

Proof of Proposition A1. Suppose (x_1, y_1) and (x_2, y_2) satisfy $\Phi(x_1, y_1) = \Phi(x_2, y_2)$. From the equalities of j and r , we have

$$x_1 \equiv x_2 \pmod{2}, \quad \left\lfloor \frac{x_1}{2} \right\rfloor = \left\lfloor \frac{x_2}{2} \right\rfloor. \quad (\text{A2})$$

Since any nonnegative integer x is uniquely determined by its parity $x \bmod 2$ and its quotient $\lfloor x/2 \rfloor$, it follows that $x_1 = x_2$.

From the equality of q , we have $y_1 \equiv y_2 \pmod{2}$. From the equality of m , together with $x_1 = x_2$, we obtain

$$\left\lceil \frac{x_1 + y_1}{2} \right\rceil = \left\lceil \frac{x_1 + y_2}{2} \right\rceil. \quad (\text{A3})$$

Suppose for contradiction that $y_1 \neq y_2$. Since they have the same parity, $|y_1 - y_2| \geq 2$. Assume without loss of generality that $y_1 > y_2$, and let $d = (y_1 - y_2)/2 \geq 1$. Then

$$\frac{x_1 + y_1}{2} = \frac{x_1 + y_2}{2} + d. \quad (\text{A4})$$

Since d is a positive integer, applying the ceiling function yields

$$\left\lceil \frac{x_1 + y_1}{2} \right\rceil = \left\lceil \frac{x_1 + y_2}{2} \right\rceil + d > \left\lceil \frac{x_1 + y_2}{2} \right\rceil, \quad (\text{A5})$$

which contradicts Equation (A3). Therefore $y_1 = y_2$. Hence $(x_1, y_1) = (x_2, y_2)$. $\square$

Proposition A1 establishes that the refined quadruple (m, j, r, q) carries the same information as the physical state (x, y) : distinct physical states never map to the same quadruple.

Appendix A.2. Equivalence of Equilibrium Thresholds

We now prove that replacing the aggregated mapping used in the main text with the refined mapping Φ does not alter the threshold strategy.

The expected payoff of a customer who observes state $(m, 0)$ is given by $U_0(m)$ in Equation (19), and the expected payoff of a customer who observes state $(m, 1)$ is given by $U_1(m)$ in Equation (15). Both payoff functions depend only on the total number of batches m , not on the number of complete batches in the queue r or on the indicator q of whether a single customer is in service. In other words, any two physical states sharing the same m and the same j yield identical expected payoffs, regardless of their r and q values.

Consequently, the equilibrium thresholds in the main text,

$$m_0^e = \max\{m > 0 : U_0(m) \geq 0\}, \quad m_1^e = \max\{m > 0 : U_1(m) \geq 0\}, \quad (\text{A6})$$

are determined solely by the payoff functions $U_0(m)$ and $U_1(m)$, and are independent of r and q .

Proposition A2. *The equilibrium thresholds obtained from the aggregated mapping $(x, y) \rightarrow (m, j)$ and from the refined mapping Φ are identical.*

Proof of Proposition A2. The equilibrium threshold m_0^e is defined as the largest integer satisfying $U_0(m) \geq 0$, and m_1^e as the largest integer satisfying $U_1(m) \geq 0$. The payoff functions $U_0(m)$ and $U_1(m)$ have exactly the same expressions under both the aggregated and refined mappings, since they depend only on m , not on r or q . Therefore, the thresholds are identical under both mappings, and consequently m_0^e and m_1^e coincide. $\square$

Proposition A2 confirms that the aggregated state (m, j) used in the main text is sufficient for the equilibrium analysis. Although the refined quadruple provides a more complete description of the system state, it does not alter any analytical results, including the threshold strategies, the steady-state probability distribution, or the performance measures derived in Section 4.

References

1. Naor, P. The regulation of queue size by levying tolls. *Econometrica* **1969**, *37*, 15–24. <https://doi.org/10.2307/1909200>.
2. Edelson, N.M.; Hilderbrand, D.K. Congestion tolls for Poisson queuing processes. *Econometrica* **1975**, *43*, 81–92. <https://doi.org/10.2307/1913415>.
3. Xu, F.; Tian, R.L.; Shao, Q. Optimal joining strategy and pricing analysis in unreliable retrial queues with predictive maintenance. *RAIRO-Oper. Res.* **2026**, *60*, 55–77. <https://doi.org/10.1051/ro/2025150>.
4. Sylia, A.; Samira, T. On a Balking M/G/1 Queue with General Differentiated Vacations. *Methodol. Comput. Appl. Probab.* **2025**, *27*, 18. <https://doi.org/10.1007/s11009-025-10145-x>.
5. Snitkovsky, R.; Roet-Green, R.; Ji, J.W. Foresee the next line: customer strategies and information disclosure in tandem queues. *Oper. Res.* **2025**, *74*, 682–699. <https://doi.org/10.1287/opre.2023.0320>.
6. Chauhan, R.S.; Selvaraju, N. Strategic behavior and social optimization in Markovian working vacation queues with delayed observations. *Oper. Res.* **2025**, *25*, 52. <https://doi.org/10.1007/s12351-025-00919-z>.
7. Varshney, S.; Acharya, M.; Kanjibhai, Z.J.; Pushkarna, M.; El-Shafai, W.; Zaitsev, I. Economic and Sensitivity Analyses of Controllable Queues With Customer Feedback and Impatience Using Markovian Modeling. *J. Math.* **2025**, *2025*, 9648114. <https://doi.org/10.1155/jom/9648114>.
8. Song, Y.; Liu, X.Y. Optimal strategy and system performance in a feedback queue with customer-intensive services and boundedly rational customers. *Methodol. Comput. Appl. Probab.* **2025**, *27*, 17. <https://doi.org/10.1007/s11009-025-10144-y>.

9. Zhu, S.; Han, X.G.; Wang, J.T.; Wang, Z.B. Performance evaluation and equilibrium strategies for a mixed priority queueing system. *Queueing Syst.* **2026**, *110*, 20. <https://doi.org/10.1007/s11134-026-09982-5>.
10. Tian, R.L.; Wang, T.; Su, J.T. Strategies analysis and ANFIS computing of the M/M/1 retrial queue with vacations and preemptive priority. *Commun. Stat.-Theory Methods* **2026**, *55*, 3183–3204. <https://doi.org/10.1080/03610926.2025.2568716>.
11. Bountali, O.; Economou, A. Equilibrium joining strategies in batch service queueing systems. *Eur. J. Oper. Res.* **2017**, *260*, 1142–1151. <https://doi.org/10.1016/j.ejor.2017.01.024>.
12. Nguyen, H.Q.; Phung-Duc, T. To wait or not to wait: Strategic behaviors in an observable batch-service queueing system. *Oper. Res. Lett.* **2022**, *50*, 343–346. <https://doi.org/10.1016/j.orl.2022.04.003>.
13. Nakamura, A.; Phung-Duc, T. Equilibrium analysis for batch service queueing systems with strategic choice of batch size. *Mathematics* **2023**, *11*, 3956. <https://doi.org/10.3390/math11183956>.
14. Sun, W.; Xie, X.M.; Zhang, Z.Y.; Li, S.Y. Customer joining strategies in Markovian queues with B-limited service rule and multiple vacations. *4OR-Q. J. Oper. Res.* **2024**, *22*, 211–233. <https://doi.org/10.1007/s10288-023-00542-8>.
15. Liu, X.L.; Xu, X.L.; Liu, M.X. Strategy analysis of retrial queue with parallel customer and standby server. *Methodol. Comput. Appl. Probab.* **2024**, *26*, 27. <https://doi.org/10.1007/s11009-024-10098-7>.
16. Hanukov, G.; Perel, E.; Perel, N. On strategic customers in the group-merging Israeli queue. *Queueing Syst.* **2025**, *109*, 16. <https://doi.org/10.1007/s11134-025-09944-3>.
17. Fielbaum, A.; Kucharski, R.; Cats, O.; Alonso-Mora, J. How to split the costs and charge the travellers sharing a ride? aligning system's optimum with users' equilibrium. *Eur. J. Oper. Res.* **2022**, *301*, 956–973. <https://doi.org/10.1016/j.ejor.2021.11.041>.
18. Wang, Y.L.; Cao, P.; Xie, J.G.; Zhang, D.Y. Advance Selling and Upgrading in Priority Queues. *Manuf. Serv. Oper. Manag.* **2025**, *27*, 1664–1682. <https://doi.org/10.1287/msom.2023.0410>.
19. Nakamura, A.; Phung-Duc, T. Fair and efficient sharing: Dynamic pricing control for batch service system with strategic customers. *Transp. Res. Part C Emerg. Technol.* **2025**, *171*, 104994. <https://doi.org/10.1016/j.trc.2024.104994>.
20. Bergquist, J.; Elmachtoub, A.N. Static pricing guarantees for queueing systems. *Stoch. Syst.* **2026**, *16*, 1–21. <https://doi.org/10.1287/stsy.2023.0057>.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.