mathematics

MDPI

Article

Equilibrium Analysis for Batch Service Queueing Systems with Strategic Choice of Batch Size

Ayane Nakamura [1,†] and Tuan Phung-Duc [2,*,†]

[1] Graduate School of Science and Technology, University of Tsukuba, Tsukuba 305-8573, Japan; s2230122@u.tsukuba.ac.jp
[2] Institute of Systems and Information Engineering, University of Tsukuba, Tsukuba 305-8573, Japan
* Correspondence: tuan@sk.tsukuba.ac.jp
† These authors contributed equally to this work.

Abstract: Various transportation services exist, such as ride-sharing or shared taxis, in which customers receive services in a batch of flexible sizes and share fees. In this study, we conducted an equilibrium analysis of a variable batch service model in which customers who observe no waiting customers in an incomplete batch can strategically select a batch size to maximize the individual utilities. We formulated this model as a three-dimensional Markov chain and created a book-type transition diagram. To consider the joining/balking dilemma of customers for this model, we proposed an effective algorithm to construct a necessary and sufficient size of state space for the Markov chain provided that all customers adopt the threshold-type equilibrium strategy. Moreover, we proved that the best batch size is a non-decreasing function for i if the reward for the completion of batch service with size l is an increasing function of l assuming that a tagged customer observes i complete batches in the system upon arrival; in other words, the fee decreases as the batch becomes larger. We then derive several performance measures, such as throughput, social welfare, and monopolist's revenue. Throughout the numerical experiment, a comparison between the present variable batch service model and regular batch service model in which customers were served in a constant batch, was discussed. It was demonstrated that the three performance measures can be optimized simultaneously in the variable batch service model, as long as the fee was set relatively high.

Keywords: queues with strategic customers; nash equilibrium; batch service; variable batch size; strategic choice of batch size; social optimization; revenue management

MSC: 90B22; 60K25; 60K30

Citation: 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

Academic Editor: José Niño-Mora

Received: 28 August 2023
Revised: 14 September 2023
Accepted: 14 September 2023
Published: 18 September 2023

1. Introduction

Batch service queueing systems have frequently appeared in recent transportation services. Examples are shared taxis or ride-sharing services, such as Uber or Lyft. In these systems, customers typically form a group, ride in the same vehicle, and share a move. Then, the concern over the 'best' batch size has risen. Customers generally share fees with other customers in a batch; therefore, the financial burden on a customer decreases as the batch size increases. However, customers must experience longer waiting times to create a larger batch, that is, to wait for the arrival of other customers. Therefore, to determine the ideal batch size, constructing a model that considers both financial and time costs is necessary.

In the field of queueing theory, many studies that consider the behavior of strategic customers have been conducted to discuss systems from an economic viewpoint. Regarding batch service queueing systems, some studies have been conducted on models in which strategic customers are served in a batch with a constant batch size. However, to the best of our knowledge, a model in which customers strategically select a batch size to maximize their individual utility based on the state of the system has never been studied. In this

study, we analyzed such a dynamic batch service queueing system with strategic customers (see the detailed explanation in Section 2). From a managerial perspective, for modern transportation systems, a scope for discussion regarding the effect of the strategic choice of batch size on both social welfare and administrator revenue still exists. In addition, the analysis of this model is challenging from a modeling perspective. The arrival interval of complete batches to the system depends on the congestion level in the present model, whereas the arrival process in the constant batch size model can be expressed by the Erlang distribution with a rate equivalent to the arrival rate of customers and shape, which is determined by the batch size.

The batch service queueing model was pioneered by Bailey [1], and many studies have been conducted for more than half a century. We refer to the thorough survey in [2] and only cite the research that handles a non-constant batch size:

- General bulk-service (GBS) queues. The GBS rule (or (a, b)-rule) is explained as follows. The server will start to provide service only when at least 'a' customers are in the queue, and the maximum service capacity is 'b'. This type of queue has been studied extensively (see, e.g., [3–13]). As finite buffer models, a few studies have been conducted (see, e.g., [14–19]). In addition, the models of batch-size-dependent service have been studied [13,18].
- Batch service queue with a random batch size. The models where the size of a batch service follows an arbitrary distribution have been conducted [20,21]. In addition, the model of batch-size-dependent service has been studied [22].

Queues with strategic customers were first investigated by Naor [23], who studied the joining/balking dilemma in an observable M/M/1 queue. Many studies have been conducted on the game-theoretical analysis of queues (see surveys [24,25]). Previous studies on batch service queueing systems can be summarized as follows:

- Regular batch service. Several studies have considered the system in which the service is conducted by a batch of constant batch size. The observable and unobservable models [26] as well as the partially observable models [27,28] have been studied. A simple extension to the multiple server model of the observable case has also been conducted [29]. Furthermore, an equilibrium analysis of a batch service queueing system considering the limitation for the number of service facilities (e.g., cars in the application of real life) has been studied [30].
- Clearing system. Some research has been conducted on the clearing system in which the service facilities periodically remove all present customers. The clearing system with strategic customers was first studied in [31], which assumes an alternative environment scheme. Moreover, a clearing system with bounded rationality was considered [32]. Furthermore, many policies of information disclosure levels were compared (see, e.g., [33–35]). The analysis of a transportation system with constant intervisit time and heterogeneous customers was also conducted [36].
- Choice among multiple capacity services. Several studies have been conducted on the strategic choice among infinite server systems with regular batches and single-server systems of single service [37–41]. These studies showed interesting Downs–Thompson and Braess-type paradoxes through the equilibrium analyses of this type of model. Some recent studies considered the routing problems in a parallel clearing system [42] and the strategical choice between a single service and batch service with size two on the basis of waiting time cost [43].

This study is related to pricing control among different types of services. The aforementioned study on clearing systems [34] also considered the pricing scheme. Additionally, we cite the following papers:

- Priority queues. One fundamental method of pricing control is the introduction of priorities. In priority queues, customers can buy the priorities with the additional fee and replace the ordinary (not buying priority) customers, saving the waiting time (see,

e.g., Chapter 4 in [24] and [44,45]). Both social optimization and revenue maximization were considered in [44].

In this study, we proposed a new type of observable batch service queueing system. This mechanism can be briefly explained as follows: If no complete batch exists when a new customer arrives in the system, he can decide the size of his batch to be served (or balking the system) based on the expected utility, which depends on the expected sojourn time and fee. A new customer, who finds any waiting customers in an incomplete batch, decides to join the batch or balk the system based on expected utility. This model corresponds to the recent shared-taxi or ride-sharing system, where customers create a group of favorable size based on the shared fee and the expected time cost, which is influenced by the arrival rate of customers and the congestion state of the road.

From a technical perspective, we extended the two-dimensional Markov chain in the regular batch service model [26] to a three-dimensional Markov chain that considers the batch size of an incomplete batch. We found that our Markov chain forms the 'book-type' state space. By leveraging this form, we calculated the expected sojourn time for a tagged customer. To consider the joining/balking dilemma of customers for this model, we proposed an effective algorithm to construct a necessary and sufficient size of state space for the Markov chain provided that all customers adopt the threshold-type equilibrium strategy. Furthermore, we proved that the best batch size (for the first-arriving customer who observes zero waiting customers in an incomplete batch), provided that a tagged customer observes i complete batches in the system upon arrival, is a non-decreasing function for i if the reward for the completion of a batch service with size l is an increasing function of l; in other words, the fee decreases as the batch becomes larger.

We conducted various numerical experiments on throughput, social welfare, and revenue. In the experiment, we introduced the notions of 'fixed fee' and 'sharing fee'. A fixed fee refers to the fee that all customers pay to receive the service, regardless of their batch size. The sharing fee refers to the fee for a batch, that is, the fee for a person is the sharing fee divided by the batch size. In this setting, we observed customer behavior by considering the trade-off between time and money.

The proposed model encompasses Naor's model [23] and the observable regular batch service model [26]. Throughout the numerical experiment, we demonstrated some performance measures: throughput, social welfare, monopolist's revenue, and expected sojourn time in the variable batch service model show more stable performance than the regular batch service model if both sharing and fixed fees are set to some extent.

In summary, the important messages of this study from the application point of view are as follows: the system where the administrator enables customers to form the favorable size of the batch to maximize their utility is meaningful, particularly when the fixed and sharing fees are relatively high. The proposed model does not show extremely bad results for any arrival rate as long as both fees are set in a well-balanced manner compared to the regular batch size model. This indicates that the proposed model is a robust model for the fluctuation of customer demand. In addition, the introduction of the variable batch service model can lead to optimizing all of the throughput, social welfare, and monopolist's revenue, simultaneously, compared to the regular batch service model under a high setting of the fees. On the other hand, the variable batch service model shows bad performance when the fees are set low.

The remainder of this paper is organized as follows. The settings of the model are described in detail in Section 2. Section 3 presents the analysis of the model. We present several numerical examples in Section 4. Finally, Section 5 concludes the paper.

2. Modeling

This section describes the model settings in detail. Customers arrive at the queue according to a Poisson process at a rate λ one-by-one and receive the service as a batch. The (single) server service time for a batch follows an exponential distribution of the parameter μ regardless of the batch size. The reward for receiving a service is defined by R and the

customer fee for a service with batch size l is denoted by p_l. Therefore, the reward for receiving a service in a batch of size l can be regarded as $R_l = R - p_l$. Furthermore, R_l becomes a function of l with upper bound R. We assume that all customers are rational and have a common knowledge of the aforementioned information. Moreover, we assume that the system is observable; that is, customers can observe the number of complete batches in the system, the number of waiting customers in an incomplete batch, and the size of the incomplete batch (when it will become full) at the time of their arrival.

If a customer finds no waiting customers in the incomplete batch (i.e., all batches in the system are complete batches and he cannot join them) and i complete batches in the system, he decides to adopt the best batch size l^* according to the following definition:

$$l^* = \begin{cases} 0, & \text{if } R_l - CW_{i,0,l} < 0 \text{ for } \forall l, \\ \max \mathbb{D}, & \text{otherwise,} \end{cases} \tag{1}$$

and

$$\mathbb{D} = \underset{l \in \mathbb{N}}{\arg\max} \, (R_l - CW_{i,0,l}),$$

where $W_{i,0,l}$ denotes the expected sojourn time for a tagged customer who observes zero waiting customers in an incomplete batch, under the condition that the number of complete batches is i at the arrival time point, and the customer adopts batch size l. Here, note that $\max \mathbb{D}$ stands for the maximum element (batch size) in set $\mathbb{D}$. If no incentive exists to join the system for any batch size, we let l^* be 0.

If a tagged customer finds i complete batches, k (>0) customer(s) in the incomplete batch, and the batch size that he receives from the service is l at his arrival time point, they join the system with probability $q_{i,k,c}$ as follows:

$$q_{i,k,l} = \begin{cases} 1, & \text{if } R_l - CW_{i,k,l} \geq 0, \\ 0, & \text{if } R_l - CW_{i,k,l} < 0, \end{cases} \tag{2}$$

where $W_{i,k,l}$ represents the expected sojourn time (if he joins) under the condition that the number of complete batches and the number of customers in the incomplete batch are i and k, respectively, at his arrival time point with batch size of l.

We assume that not all customers are allowed to change their decisions (best batch size, join/balk) after deciding. For simplicity, we assume that customers are in favor of joining a larger batch; that is, when they are indifferent between joining and balking, they choose to join, and when they are indifferent among multiple batch sizes, they choose the largest batch, similar to that in [26].

Let $I(t)$, $K(t)$, and $L(t)$ denote the number of complete batches, number of waiting customers in the incomplete batch, and the size of the incomplete batch when it becomes full, respectively, at time t. Then, $\{(I(t), K(t), L(t)); t \geq 0\}$ becomes an irreducible three-dimensional Markov chain with state space $\mathcal{U}$, where

$$\mathcal{U} = \mathbb{Z} \times \mathcal{V},$$

$$\mathcal{V}_1 = \{(0,1)\},$$

$$\mathcal{V}_a = \{(1,a), (2,a), \ldots (a-1,a)\}, \qquad a \geqq 2,$$

$$\mathcal{V} = \{\mathcal{V}_1, \mathcal{V}_2, \mathcal{V}_3 \ldots .\}.$$

Notably, $L(t) = 1$ is a special case because $L(t)$ accepts 1 automatically whenever $K(t) = 0$ holds, because of the aforementioned definition. Here, we let I, K, L denote the random variables in the steady state, that is, the variables under $t \to \infty$.

We refer to the transition diagram ($1 \leq L \leq 3$) in Figure 1 and schematic of the overall transition structure in Figure 2. Let $q_{i,0,l}$ in the diagram denote the probability that a tagged customer who observes i complete batches and 0 customers in the incomplete batch chooses batch size l, similar to that in (2). Notably, Figure 2 shapes 'book type' form in which the

axis corresponds to the states where any incomplete batch does not exist, and the pages correspond to the other state. In other words, the axis corresponds to the special case of $(K, L) = (0, 1)$. Based on this three-dimensional Markov chain, we propose algorithmic procedures for analyzing the model in Section 3.

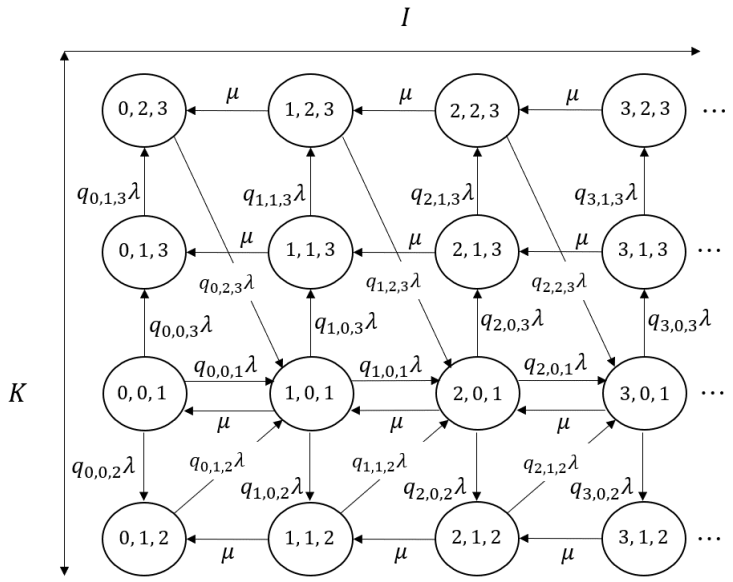


Figure 1. The transition diagram ($1 \leq L \leq 3$).

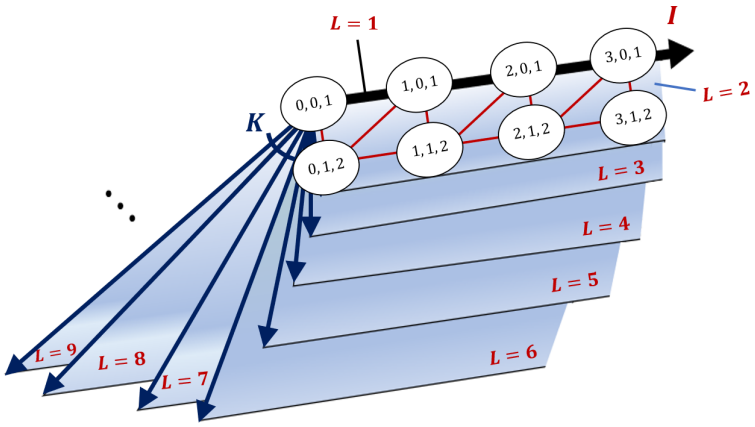


Figure 2. Schematic of the transition diagram.

The proposed model encompasses Naor's model [23] (observable M/M/1 model), where we set $R_1 > 0$ and $R_l = 0$ for $\forall l \in \mathbb{N} \setminus \{1\}$ and a regular batch service model [26] with batch size K, where we set $R_K > 0$ and $R_l = 0$ for $\forall l \in \mathbb{N} \setminus \{K\}$. A comparison of these three models is presented in Section 4.

The application of this model is as follows. There are sufficient cars at the boarding station on the transportation platform, such as ride-sharing or shared taxis. Customers with the same destination arrive at this platform according to a Poisson process. A customer who observes no waiting customers at his arrival point can decide the size of his group to share the service (or abandon the service), whereas other customers decide whether to join the collecting group. When a group is completed, the customers in the group ride the car and instantly depart from the boarding station. The car then joins a 'service station' [46] on the road. A service station is a single-server queue representing a road segment. We refer to this method to express traffic congestion on roads in terms of the queueing model proposed in [46]. The proposed model enables us to study strategic customers' behavior by considering the trade-off between waiting time, traffic congestion, and cost.

3. Analysis

This section presents an analysis of the proposed model. We change the notation of l^* in (1) in Section 2 to $l^*(i)$ to emphasize the dependency on i. In addition, we refer $l^*(i)$ to as the 'best batch size' when a tagged customer observes i complete batches. Furthermore, we introduce several notations. Let $i_{(\max)}$ denote the maximum i such that it satisfies $l^*(i) > 0$; that is,

$$i_{(\max)} = \max\,(i \mid l^*(i) > 0).$$

In addition, we let $i_{(l*\max)}$ and $i_{(l*\min)}$ denote the maximum and minimum i, respectively, such that they satisfy $l^*(i) = l\ (> 0)$; that is,

$$i_{(l*\max)} = \max\,(i \mid l^*(i) = l),$$

$$i_{(l*\min)} = \min\,(i \mid l^*(i) = l).$$

We also define a set $\mathbb{B}_{(i)}$ that stands for the set of all possible best batch sizes where $i \le i' \le i_{(\max)}$ as follows:

$$\mathbb{B}_{(i)} = \bigcup_{i \le i' \le i_{(\max)}} \{l^*(i')\}.$$

One technical point in the analysis of this model is the derivation of the expected sojourn time for a tagged customer, which is related to his decision. As described in Section 2, this model constitutes a three-dimensional Markov chain; therefore, deriving the expected sojourn time depending on the three simultaneous states is necessary.

This can be calculated using a recursive procedure similar to that of the regular batch service model [26] by leveraging the book type form of the transition diagram in Figure 2; the states, wherein the number of customers in the incomplete batches is 0, are depicted as the black axis, and other states (i.e., some waiting customers exist in an incomplete batch) are expressed as pages that correspond to the batch sizes, and each page is connected to the axis. We can then easily determine that the transition on each page is equivalent to a regular batch service model [26] except for that each arrival rate λ is multiplied by the joining probability $q_{i,k,l}$.

Lemmas 1 and 2 are implied from Theorems 4.1 and 4.2 in [26]. Here for the readability, we rewrite them and their proofs in our notations. We show Lemma 1 as follows:

Lemma 1. *A unique equilibrium strategy* $(q^e_{i,k,l}; (i,k,l) \in \mathcal{U})$ *exists, as*

$$q^e_{i,0,l} = \begin{cases} 1, & \text{if } l = l^*(i), \\ 0, & \text{if } l \neq l^*(i), \end{cases} \quad k = 0, \tag{3}$$

$$q^e_{i,k,l} = \begin{cases} 1, & \text{if } i \leqq i^e_{k,l}, \\ 0, & \text{if } i > i^e_{k,l}, \end{cases} \quad 1 \le k \le l-1, \tag{4}$$

where $i^e_{1,l} \leqq i^e_{2,l} \leqq \cdots \leqq i^e_{l-1,l}$.

Proof. If an arriving customer who observes state $(i, l-1, l)$ joins the system, his sojourn time clearly becomes the sum of $i+1$ exponential distributions with parameter μ. This is because his decision is dominant; that is, his decision is not affected by the decisions of other customers. Therefore, we obtain

$$q^e_{i,l-1,l} = \begin{cases} 1, & \text{if } i \leqq i^e_{l-1,l}, \\ 0, & \text{if } i > i^e_{l-1,l}, \end{cases}$$

where

$$i^e_{l-1,l} = \left\lfloor \frac{\mu R_l}{C} - 1 \right\rfloor.$$

Subsequently, we consider an arriving customer who observes the state $(i, l-2, l)$. Clearly, $q^e_{i,l-2,l} = 0$ if $q^e_{i,l-1,l} = 0$. This is because the expected sojourn time of an arriving customer who observes state $(i, l-2, l)$ is longer than that of an arriving customer who observes state $(i, l-1, l)$ because the customer must wait for the next customer to arrive. In addition, we find that $W_{i,k,l}$ is an increasing function of i and tends to ∞ as $i \to \infty$ (as we will describe in Lemma 3). Therefore, a unique $i^e_{l-2,l}$ exists such that $W_{i^e_{l-2,l},k,l} \leq R_l/C < W_{i^e_{l-2,l}+1,k,l}$. Based on the aforementioned discussion, we obtain the equilibrium joining probability as follows:

$$q^e_{i,l-2,l} = \begin{cases} 1, & \text{if } i \leqq i^e_{l-2,l}, \\ 0, & \text{if } i > i^e_{l-2,l}, \end{cases}$$

where

$$i^e_{l-2,l} \leq i^e_{l-1,l}.$$

By repeating this process for $k = l-3, l-4$, and... in the descending order, we obtain (4). The case $k = 0$ in (3) is notable because of the definition of (1). □

To calculate the threshold $i^e_{k,l}$, we must derive the expected sojourn time conditional on each state. The procedure for deriving the expected sojourn time is provided in Lemma 2.

Lemma 2. *$W_{i,k,l}$ for $k \geq 1$ can be calculated by using the following recursive scheme:*

$$W_{0,k,l} = \frac{l-k-1}{\lambda} + \frac{1}{\mu}, \qquad l \in \mathbb{N} \text{ and } 0 \leq k \leq l-1, \tag{5}$$

$$W_{i,l-1,l} = \frac{i+1}{\mu}, \qquad i \in \mathbb{Z} \text{ and } l \in \mathbb{N} \setminus \{1\}, \tag{6}$$

$$W_{i,k,l} = \frac{1}{\lambda+\mu} + \frac{\lambda}{\lambda+\mu} W_{i,k+1,l} + \frac{\mu}{\lambda+\mu} W_{i-1,k,l}, \qquad i \in \mathbb{N} \text{ and } l \in \mathbb{N} \text{ and } 0 \leq k \leq l-2, \tag{7}$$

or the formulae

$$W_{i,k,l} = \frac{i+1}{\mu} + \frac{1}{\lambda}\left(\frac{\mu}{\lambda+\mu}\right)^i \sum_{b=0}^{l-k-1} (l-k-b-1)\binom{i+b-1}{b}\left(\frac{\lambda}{\lambda+\mu}\right)^b, \qquad i \geq 1, \tag{8}$$

and (5).

Proof. Notably, when the batch size for the next batch is determined, it cannot be modified until the batch becomes complete. In addition, the expected sojourn time of a tagged arriving customer depends only on the number of complete batches and the number of waiting customers of an incomplete batch and is not affected by the batch sizes, except for the incomplete batch. Therefore, this lemma becomes equivalent to Theorem 4.2 in [26] when we regard l as a constant batch size K.

Specifically, (5)–(7) can be shown naturally in the first-step analysis. For (8), $W_{i,k,l}$ can be expressed as follows:

$$W_{i,k,l} = \frac{1}{\mu} + \mathrm{E}[\max\{Y_i, Z_{l-k-1}\}] = \frac{i+1}{\mu} + \mathrm{E}[\max\{0, Z_{l-k-1} - Y_i\}], \tag{9}$$

where Y_i and Z_{l-k-1} are Erlang-i and Erlang-$(l-k-1)$ independent random variables with rates μ and λ, corresponding to the total service time of the present complete batches and the completion time of the current incomplete batch, respectively, as described in [26]. Then, we obtain the following transformation:

$$\begin{aligned}
&\mathrm{E}[\max\{0, Z_{l-k-1} - Y_i\}] \\
&= \sum_{b=0}^{\infty} \mathrm{P}[N(Y_i) = b]\mathrm{E}[\max\{0, Z_{l-k-1} - Y_i\}|N(Y_i) = b] \\
&= \sum_{b=0}^{l-k-1} \mathrm{P}[N(Y_i) = b]\frac{l-k-b-1}{\lambda},
\end{aligned} \tag{10}$$

where $N(Y_i)$ represents the number of Poisson events with a rate λ during Erlang$i$ time, that is, Y_i with rate μ. We can observe that

$$\mathrm{P}[N(Y_i) = b] = \left(\frac{\mu}{\lambda+\mu}\right)^i \binom{i+b-1}{b} \left(\frac{\lambda}{\lambda+\mu}\right)^b \tag{11}$$

holds based on tedious calculations. □

Based on Lemma 2, the following lemmas hold:

Lemma 3. *Regarding $W_{i,k,l}$ in Lemma 2, the following properties hold:*

(i) $W_{i,k,l}$ *is a strictly increasing function of* i.
(ii) $W_{i,k,l}$ *is a strictly decreasing function of* k.
(iii) $W_{i,k,l}$ *is a strictly increasing function of* l.
(iv) $\lim_{i\to\infty} W_{i,k,l} = \infty$.
(v) $\lim_{l\to\infty} W_{i,k,l} = \infty$.
(vi) $W_{i,k,l} - W_{i,k,l'}$ *is a strictly decreasing function of* i *for* $\forall l' < l$.

Proof. In this lemma, (i), (ii), and (iv) correspond to (i)–(iii) in Theorem 4.2 in [26]. As shown, (ii)–(v) hold true for (8). The definition in (9) immediately yields (i). Regarding (vi), we obtain the following relationship from (9)–(11):

$$\begin{aligned}
&W_{i,k,l} - W_{i,k,l-1} \\
&= \frac{1}{\lambda} \sum_{b=0}^{l-k-2} \left(\frac{\mu}{\lambda+\mu}\right)^i \binom{i+b-1}{b} \left(\frac{\lambda}{\lambda+\mu}\right)^b \\
&= \frac{1}{\lambda}\mathrm{P}[N(Y_i) \le l-k-2].
\end{aligned} \tag{12}$$

Owing to the definition of $N(Y_i)$ in the proof of Lemma 2, we show that $W_{i,k,l} - W_{i,k,l-1}$ is a strictly decreasing function of i for any l, which supports (vi). □

Lemma 4. *The thresholds $i^e_{k,l}$ $(k > 0)$ in Lemma 1 are*

$$i^e_{l-1,l} = \left\lfloor \frac{\mu R_l}{C} - 1 \right\rfloor,$$

$$i^e_{k,l} = \max\left\{ i; 0 \leqq i \leqq i^e_{k+1,l} \text{ and } W_{i,k,l} \leqq \frac{R_l}{C} \right\}, \qquad 1 \le k \le l-2.$$

Proof. The monotonicity of $W_{i,k,l}$ for i–that is, (i) in Lemma 3 yields the result. □

Next, we derive the best batch size, that is, $l^*(i)$. We show Lemma 5.

Lemma 5. $l^*(i) = l$ *holds if* $i_{(l*min)} < i < i_{(l*max)}$.

Proof. For any l such that $l > \tilde{l}$, $R_l - CW_{i_{(l*\min)},0,l} \geq R_{\tilde{l}} - CW_{i_{(l*\min)},0,\tilde{l}}$ and $R_l - CW_{i_{(l*\max)},0,l} \geq R_{\tilde{l}} - CW_{i_{(l*\max)},0,\tilde{l}}$, that is, $(R_l - R_{\tilde{l}})/C \geq W_{i_{(l*\min)},0,l} - W_{i_{(l*\min)},0,\tilde{l}}$ and $(R_l - R_{\tilde{l}})/C \geq W_{i_{(l*\max)},0,l} - W_{i_{(l*\max)},0,\tilde{l}}$ clearly hold from the definition. Considering the

monotonicity of $W_{i,k,l} - W_{i,k,l-1}$ with respect to i ((vi) in Lemma 3), it is clear for $i_{(l*\min)} < i < i_{(l*\max)}$ that $(R_l - R_{\tilde{l}})/C \geq W_{i,0,l} - W_{i,0,\tilde{l}}$ holds, i.e., $R_l - CW_{i,0,l} \geq R_{\tilde{l}} - CW_{i,0,\tilde{l}}$ holds.

Similarly, for any l such that $l < \tilde{l}$, $R_l - CW_{i_{(l*\min)},0,l} > R_{\tilde{l}} - CW_{i_{(l*\min)},0,\tilde{l}}$ and $R_l - CW_{i_{(l*\max)},0,l} > R_{\tilde{l}} - CW_{i_{(l*\max)},0,\tilde{l}}$ that is, $W_{i_{(l*\min)},0,\tilde{l}} - W_{i_{(l*\min)},0,l} > (R_l - R_{\tilde{l}})/C$ and $W_{i_{(l*\max)},0,\tilde{l}} - W_{i_{(l*\max)},0,l} > (R_l - R_{\tilde{l}})/C$ clearly hold from the definition. Therefore, based on the monotonicity of (vi) in Lemma 3, $R_l - CW_{i,0,l} > R_{\tilde{l}} - CW_{i,0,\tilde{l}}$ holds true for $i_{(l*\min)} < i < i_{(l*\max)}$.

By combining the aforementioned discussions, we obtain Lemma 5. This lemma becomes important for the proofs of Theorem 1 and Lemma 6. □

Based on Lemma 5, Theorem 1 can be shown as follows:

Theorem 1. *$\mathbb{B}_{(i)}, i_{(\max)}, i_{(l*\min)}$, and $i_{(l*\max)}$ can be calculated by Algorithm 1.*

Algorithm 1: Calculate $\mathbb{B}_{(i)}, i_{(\max)}, i_{(l*\min)}, i_{(l*\max)}$.

Input: $\lambda, \mu, R_l, R, C, W_{i,k,l}$ (from Lemma 2)
Output: $\mathbb{B}_{(i)}, i_{(\max)}, i_{(l*\min)}, i_{(l*\max)}$
calculate $l^*(i), i_{(l*\min)}, i_{(l*\max)}, i_{(l*\max)}$:
1: **for** $i = 0, 1, \cdots$ **do**
2: ## start from $l = 1$:
3: **while** $R > CW_{i,0,l}$ **do**
4: $diff(l) \leftarrow R_l - CW_{i,0,l}$
5: $l = l + 1$
6: **end while**
7: **if** $\max\{diff(l)\} \geq 0$ **then**
8: $l^*(i) = \max\left\{ \arg\max_l \{diff(l)\} \right\}$
9: **if** $i = 0$ **then**
10: $i_{(l^*(0)*\min)} \leftarrow 0$
11: **else if** $l^*(i) \neq l^*(i-1)$ **then**
12: $i_{(l^*(i)*\min)} \leftarrow i$
13: $i_{(l^*(i-1)*\max)} \leftarrow i - 1$
14: **end if**
15: **else**
16: $l^*(i) = 0$
17: $i_{(\max)} \leftarrow i - 1$
18: $i_{(l^*(i-1)*\max)} \leftarrow i - 1$
19: break
20: **end if**
21: **end for**
calculate $\mathbb{B}_{(i)}$:
22: $\mathbb{B}_{(i_{(\max)})} \leftarrow \{l^*(i_{(\max)})\}$
23: **for** $i = i_{(\max)} - 1, i_{(\max)} - 2, \cdots, 0$ **do**
24: $\mathbb{B}_{(i)} \leftarrow \mathbb{B}_{(i+1)} \cup \{l^*(i)\}$
25: **end for**

Proof. A brief explanation of Algorithm 1 is as follows: From lines 1 to 21, $l^*(i), i_{(\max)}$, $i_{(l*\min)}$, and $i_{(l*\max)}$ are calculated in ascending order from $i = 0$. First, in the while statement in lines 3 to 6, the values of $R_l - CW_{i,0,l}$ for each l are memorized. Notably, $CW_{i,0,l}$ definitely exceeds the upper bound R in some i because of (v) in Lemma 5; therefore, the while statement is guaranteed to finish in a finite number of times. Subsequently, the best batch size, conditional on i is determined by the definition of (1) in lines 8 and 16.

The minimum i with the best batch size is $l^*(0)$; that is, $i_{(l^*(0)*\min)}$ accepts 0 (as in line 10). Whenever $l^*(i) \neq l^*(i-1)$ holds, $i_{(l^*(i)*\min)}$ and $i_{(l^*(i-1)*\max)}$ are determined (lines 12, 13, and 18) by Lemma 5, which guarantees that discontinuous i do not accept the same best batch size. The maximum i that satisfies $l^*(i) > 0$, $i_{(\max)}$, is calculated when $l^*(i)$ reaches 0, as shown in line 17. Finally, from lines 22 to 25, the set of all the possible best batches where $i \leq i' \leq i_{(\max)}$, i.e., $\mathbb{B}_{(i)}$, is determined in descending order from $i = i_{(\max)}$. □

Additionally, for the best batch sizes conditional on i, the following lemma holds:

Lemma 6. *If R_l is a strictly increasing function of l, $l^*(i)$ becomes a non-decreasing function of i (except when $l^*(i) = 0$, that is, the customer balks).*

Proof. This fact is demonstrated based on (vi) in Lemmas 3 and 5. It follows from the definition that $R_{\hat{l}} - CW_{i_{(\hat{l}*\max)},0,\hat{l}} > R_l - CW_{i_{(\hat{l}*\max)},0,l}$, evidently, $(R_{\hat{l}} - R_l)/C > (W_{i_{(\hat{l}*\max)},0,\hat{l}} - W_{i_{(\hat{l}*\max)},0,l})$ holds for $\forall l < \hat{l}$. Considering $R_{\hat{l}} - R_l$ is positive because of the assumption that R_l is a strictly increasing function of l, $(W_{i,0,\hat{l}} - W_{i,0,l})/(R_{\hat{l}} - R_l)$ is also a strictly decreasing function of i. Therefore, $1/C > (W_{i,0,\hat{l}} - W_{i,0,l})/(R_{\hat{l}} - R_l)$ always holds for $\forall i > i_{(\hat{l}*\max)}$ (C is a constant). Therefore, $\forall l < \hat{l}$ is never the best batch size, conditional on $\forall i > i_{(\hat{l}*\max)}$. Additionally, $l^*(i) = \hat{l}$ holds if $i_{(\hat{l}*\min)} \leq i \leq i_{(\hat{l}*\max)}$ from Lemma 5. Therefore, this lemma holds. This lemma is mentioned in the numerical experiment part in Section 4. □

In general shared-type transportation, customers in the same vehicle (batch) share the fee; therefore, it is natural that R_l is a strictly increasing function (i.e., p_l is a strictly decreasing function). Lemma 6 suggests that it is individually optimal for customers to make larger batches as they find a larger number of complete batches upon arrival. Lemma 6 will be mentioned in the numerical experiment part, in Remark 1, and the intuitive insight for this result in connection with the shared-type transportation examples will be described.

Based on the outputs from Algorithm 1, we can reconstruct the Markov chain for this model with reduced state space as the following theorem:

Theorem 2. *Provided that all customers adopt the threshold-type equilibrium strategy in Lemma 4, the Markov chain of our model can be defined under the reduced state space $\tilde{\mathcal{U}}$, where*

$$\tilde{\mathcal{U}} = \bigcup_{i=0}^{i_{(\max)}+1} \tilde{\mathcal{B}}_i,$$

$$\tilde{\mathcal{B}}_i = \tilde{\mathcal{V}}_{i,1} \cup \bigcup_{a \in \mathbb{B}_{(i)} \setminus \{1\}} \tilde{\mathcal{V}}_{i,a}, \qquad 0 \leq i \leq i_{(\max)},$$

$$\tilde{\mathcal{B}}_{i_{(\max)}+1} = \{(i_{(\max)}+1, 0, 1)\},$$

$$\tilde{\mathcal{V}}_{i,1} = \{(i, 0, 1)\},$$

$$\tilde{\mathcal{V}}_{i,a} = \{(i, 1, a), (i, 2, a), \ldots (i, a-1, a)\}, \qquad a \geqq 2,$$

and the transition rate from the state (i,k,l) to (i',k',l'), which is denoted by $\gamma_{(i,k,l),(i',k',l')}$, is defined as

$$\gamma_{(i,0,1),(i,1,l)} = \lambda, \qquad l \in \mathbb{B}_{(0)}, i_{(l*\min)} \leq i \leq i_{(l*\max)},$$

$$\gamma_{(i,k,l),(i,k+1,l)} = \lambda, \qquad l \in \mathbb{B}_{(i)} \setminus \{1\}, 1 \leq k \leq l-2,$$

$$\gamma_{(i,l-1,l),(i+1,0,1)} = \lambda, \qquad l \in \mathbb{B}_{(i)} \setminus \{1\},$$

$$\gamma_{(i,k,l),(i-1,k,l)} = \mu, \qquad (i,k,l) \in \tilde{\mathcal{U}} \setminus \tilde{\mathcal{B}}_0.$$

Proof. We can construct the reduced state space, $\tilde{\mathcal{U}}$, by using the results of $\mathbb{B}_{(i)}, i_{(\max)}, i_{(l*\min)}$, and $i_{(l*\max)}$ from Algorithm 1. Once a batch is completed (the capacity is satisfied), the batch has to stay in the system only for the length of Erlang distribution of rate μ and shape which is determined by the number of complete batches in the system at the moment the batch is completed (since any interruption is not allowed in this model). Hence, the state (i,k,l) such that $l \notin \mathbb{B}_{(i)}, k \geq 1$ and the state (i,k,l) for $i \geq i_{(l*\max)} + 2$ are never reached; therefore, we can remove them from the analysis. It should be noted that the state $(i,0,1)$, i.e., $\tilde{\mathcal{V}}_{i,1}$, is a special state since $l = 1$ definitely holds if $k = 0$ holds because of the aforementioned definition.

The definition of $\gamma_{(i,k,l),(i',k',l')}$ can be explained as follows. The set $\mathbb{B}_{(0)}$ stands for the set of all the possible best batch sizes within $0 \leq i \leq i_{(\max)}$ from the definition. The transition $(i,0,1) \to (i,1,l)$, i.e., the arrival of the first customer for a batch, is possible only within $i_{(l*\min)} \leq i \leq i_{(l*\max)}$, by rate λ, from Lemma 5. The arrival of customers except for first-arriving customers to the state (i,k,l) under $1 \leq k \leq l-2$ can occur for $l \in \mathbb{B}_{(i)}$. The batch completion, i.e., the transition $(i,l-1,l) \to (i+1,0,1)$, can occur as well. Finally, the service completion with rate μ occurs where $1 \leq i \leq i_{(\max)} + 1$, i.e., $(i,k,l) \in \tilde{\mathcal{U}} \setminus \tilde{\mathcal{B}}_0$. □

We here show some examples for the reconstructed transition diagrams with the state space $\tilde{\mathcal{U}}$ as follows:

Figure 3 shows the transition diagram in which $l^*(0) = 1, l^*(1) = 1, l^*(2) = 2$, $l^*(3) = 2, l^*(4) = 3$, and $l^*(5) = 0$ (thus, $i_{(\max)} = 4, i_{(1*\min)} = 0, i_{(2*\min)} = 2, i_{(3*\min)} = 4$, $i_{(1*\max)} = 1, i_{(2*\max)} = 3, i_{(3*\max)} = 4, \mathbb{B}_{(0)} = \{1,2,3\}, \mathbb{B}_{(1)} = \{1,2,3\}, \mathbb{B}_{(2)} = \{2,3\}$, $\mathbb{B}_{(3)} = \{2,3\}, \mathbb{B}_{(4)} = \{3\}$).

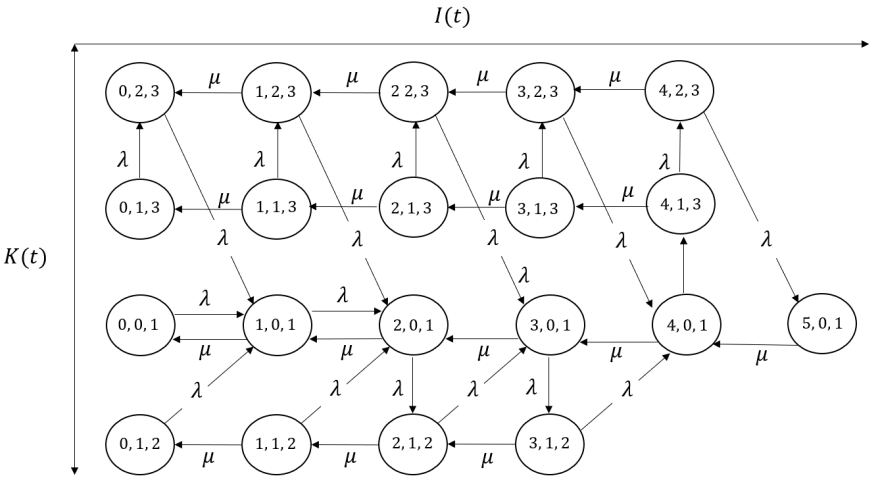


Figure 3. The reduced transition diagram for $i_{(\max)} = 4, i_{(1*\min)} = 0, i_{(2*\min)} = 2, i_{(3*\min)} = 4$, $i_{(1*\max)} = 1, i_{(2*\max)} = 3, i_{(3*\max)} = 4, \mathbb{B}_{(0)} = \{1,2,3\}, \mathbb{B}_{(1)} = \{1,2,3\}, \mathbb{B}_{(2)} = \{2,3\}, \mathbb{B}_{(3)} = \{2,3\}$, $\mathbb{B}_{(4)} = \{3\}$.

Figure 4 shows the transition diagram in which $l^*(0) = 2, l^*(1) = 3, l^*(2) = 1$, $l^*(3) = 1, l^*(4) = 1$, and $l^*(5) = 0$ (thus, $i_{(\max)} = 4, i_{(1*\min)} = 2, i_{(2*\min)} = 0, i_{(3*\min)} = 1$, $i_{(1*\max)} = 4, i_{(2*\max)} = 0, i_{(3*\max)} = 1, \mathbb{B}_{(0)} = \{1,2,3\}, \mathbb{B}_{(1)} = \{1,3\}, \mathbb{B}_{(2)} = \{1\}$, $\mathbb{B}_{(3)} = \{1\}, \mathbb{B}_{(4)} = \{1\}$).

It should be mentioned that our algorithm in Theorem 1 enables us to consider the necessary and sufficient size of the transition diagram. We can also consider the Markov chain only by deriving $i_{(\max)}$ and the maximum best batch size as well; however, by utilizing Theorem 2, we can remove the unnecessary states (i.e., all the states (i,k,l) such that $l \notin \mathbb{B}_{(i)}, k \geq 1$ are removed in $\tilde{\mathcal{U}}$), for example, $(4,1,2)$ in Figure 3 and $(1,1,2), (2,1,2), (3,1,2), (2,1,3), (2,2,3), (3,1,3), (3,2,3), (4,1,3)$ and $(4,2,3)$ in Figure 4. The number of states decreases from 21 to 20 in Figure 3 and from 21 to 11 in Figure 4. It is expected that the state space in our Markov chain becomes extreme large as $i_{(\max)}$ becomes

larger. Therefore, the calculation of $\mathbb{B}_{(i)}$ in Algorithm 1 is crucial from the perspective of the calculation cost (especially for preventing memory error). In addition, the rigorous proof in Lemma 5 also improves computational tractability. We can construct $\tilde{\mathcal{U}}$ and consider the corresponding transition rates between the states only by $\mathbb{B}_{(i)}, i_{(\max)}, i_{(l*\min)}$, and $i_{(l*\max)}$; therefore, we do not consider whether the transition $(i,0,1) \to (i,1,l)$ occurs for every i after deriving all $i_{(l*\min)}$ and $i_{(l*\max)}$. When $i_{(\max)}$ is large and the differences between $i_{(l*\min)}$ and $i_{(l*\max)}$ become larger, Lemma 5 is effective to construct $\tilde{\mathcal{U}}$ easily.

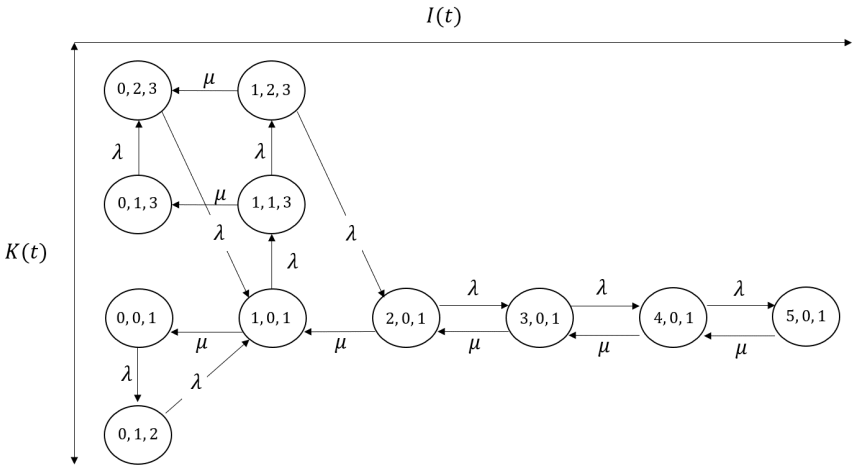


Figure 4. The reduced transition diagram for $i_{(\max)} = 4, i_{(1*\min)} = 2, i_{(2*\min)} = 0, i_{(3*\min)} = 1$, $i_{(1*\max)} = 4, i_{(2*\max)} = 0, i_{(3*\max)} = 1, \mathbb{B}_{(0)} = \{1,2,3\}, \mathbb{B}_{(1)} = \{1,3\}, \mathbb{B}_{(2)} = \{1\}, \mathbb{B}_{(3)} = \{1\}$, $\mathbb{B}_{(4)} = \{1\}$.

The Markov chain has a finite state space. Therefore, we can easily obtain the numerical values of the steady state probabilities:

$$\pi_{i,k,l} := \lim_{t\to\infty} \mathrm{P}(I(t) = i, K(t) = k, L(t) = l), \tag{13}$$

which are utilized to calculate the performance measures. First, we present Theorem 3.

Theorem 3. *Let B_{new} denote the random variables for the batch size of a newly generated batch. Then, $\mathrm{E}[B_{new}]$ and its probability function $f_{new}(n)$ are expressed by*

$$\mathrm{E}[B_{new}] = \frac{\sum_{i=0}^{i_{(max)}} \sum_{l\in\mathbb{B}_{(i)}} l\pi_{i,l-1,l}}{\sum_{i=0}^{i_{(max)}} \sum_{l\in\mathbb{B}_{(i)}} \pi_{i,l-1,l}},$$

$$f_{new}(n) = \frac{\sum_{i=0}^{i_{(max)}} \mathbb{1}_{\mathbb{B}_{(i)}}[n]\pi_{i,n-1,n}}{\sum_{i=0}^{i_{(max)}} \sum_{l\in\mathbb{B}_{(i)}} \pi_{i,l-1,l}},$$

where

$$\mathbb{1}_A[x] = \begin{cases} 1, & (x \in A), \\ 0, & (x \notin A), \end{cases} \tag{14}$$

respectively.

Proof. A new batch is generated when a customer joins the state $(K,L) = (l, l-1)$ according to the Poisson process at rate λ. Therefore, the total number of newly generated batches per unit time can be calculated as $\sum_{i=0}^{i_{(max)}} \sum_{l\in\mathbb{B}_{(i)}} \lambda\pi_{i,l-1,l}$, which is equivalent to the total rate of the transition from state $(K,L) = (l, l-1)$ to $K = 0$. Additionally, the total

number of customers served per unit time is calculated by $\sum_{i=0}^{i_{(max)}} \sum_{l\in\mathbb{B}_{(i)}} l\lambda\pi_{i,l-1,l}$. The latter divided by the former can clearly be the expected batch size. Similarly, the probability function is derived. □

In addition, we present the following theorem:

Theorem 4. *Let V denote a random variable for the number of newly generated batches per unit time. Then, $\mathrm{E}[V]$ is given by*

$$\mathrm{E}[V] = \sum_{i=0}^{i_{(max)}} \sum_{l\in\mathbb{B}_{(i)}} \lambda\pi_{i,l-1,l}.$$

Proof. The total throughput for the transition in the number of waiting customers $l-1 \to 0$ is equivalent to $\mathrm{E}[V]$. □

Notably, $\mathrm{E}[V]$ reflects traffic congestion on the road. Subsequently, we present Theorem 5.

Theorem 5. *The expected number of customers in the system $\mathrm{E}[N]$ is given as follows:*

$$\mathrm{E}[N] = \lambda\left(\sum_{i=0}^{i_{(max)}} \pi_{i,0,1}W_{i,0,l^*(i)} + \sum_{i=0}^{i_{(max)}} \sum_{l\in\mathbb{B}_{(i)}} \sum_{k=1}^{l-1} \pi_{i,k,l}W_{i,k,l}\right). \tag{15}$$

Proof. The batch size of each position is not i.i.d. distributed (because the best batch sizes are state-dependent values on i); therefore, calculating the expected number of customers is difficult in the system directly. However, the Poisson arrivals see time averages (PASTA) property holds because customers arrive at the system according to a Poisson process. We can then calculate the expected sojourn time of customers $\mathrm{E}[W]$ by conditioning $W_{i,k,l}$, using the steady state probabilities of joining customers as follows:

$$\mathrm{E}[W] = \frac{1}{\left(1-\pi_{i_{(\max)}+1,0,1}\right)}\left(\sum_{i=0}^{i_{(\max)}} \pi_{i,0,1}W_{i,0,l^*(i)} + \sum_{i=0}^{i_{(\max)}} \sum_{l\in\mathbb{B}_{(i)}} \sum_{k=1}^{l-1} \pi_{i,k,l}W_{i,k,l}\right). \tag{16}$$

By applying Little's law, the results are clear. □

Based on the aforementioned discussion, we consider several performance measures in Theorem 6.

Theorem 6. *The throughput (TH), social welfare (SW), and monopolist's revenue (MR) are given by*

$$TH = \lambda(1-\pi_{i_{(max)}+1,0,1}) = \lambda\sum_{i=0}^{i_{(max)}} \pi_{i,0,1} + \lambda\sum_{i=0}^{i_{(max)}} \sum_{l\in\mathbb{B}_{(i)}} \sum_{k=1}^{l-1} \pi_{i,k,l}, \tag{17}$$

$$SW = \lambda\sum_{i=0}^{i_{(max)}} \sum_{l\in\mathbb{B}_{(i)}} \sum_{k=1}^{l-1} \pi_{i,k,l}R + \lambda\sum_{i=0}^{i_{(max)}} \pi_{i,0,1}R - \mathrm{CE}[N], \tag{18}$$

$$MR = \lambda\sum_{i=0}^{i_{(max)}} \pi_{i,0,1}p_{l^*(i)} + \lambda\sum_{i=0}^{i_{(max)}} \sum_{l\in\mathbb{B}_{(i)}} \sum_{k=1}^{l-1} \pi_{i,k,l}p_l. \tag{19}$$

Here, TH, SW, and MR are the expected number of customers served per unit time, the difference between the generated reward and time cost per unit time, and the sum of fees that the monopolist (administrator) can obtain per unit time, respectively.

The flow of the analysis in this section can be summarized as follows. First, the equilibrium strategy of the proposed model is given by Lemma 1. Specifically, the equilibrium strategy can be calculated by using the results of Lemmas 2 and 4. Lemma 3, i.e., the properties of $W_{i,k,l}$, is used to prove Lemmas 4 and 5. Lemma 5 provides an important property of the best batch size (and it will be mentioned in Remark 6 with intuitive interpretations in the numerical experiment section). Then, Theorem 2 gives the reconstructed Markov chain, which has the necessary and sufficient number of the states, for the proposed model provided that every customer adopts the equilibrium strategy by using the outputs of Algorithm 1 in Theorem 1. Finally, some performance measures for the reconstructed Markov chain are derived in Theorems 3–6.

4. Numerical Experiments

This section presents numerical results on the performance measures in the previous section and discusses the behavior of the proposed model. Here, we introduce the notions of 'fixed fee' and 'sharing fee'. A fixed fee refers to the fee that all customers pay to receive the service, regardless of their batch size. The sharing fee refers to the fee for a batch, that is, the fee for a person is the sharing fee divided by the batch size. Let f and s denote fixed and sharing fees, respectively. In addition, the maximum batch size (corresponding to the capacity of cars in ride-sharing and shared taxis) is assumed to be m. Then, if customers belong to a batch of size l, they must pay $p_l = f + s/l$ $(1 \leq l \leq m)$. Considering these settings, the reward for a batch service of size l is given by

$$R_l = \begin{cases} R - (f + s/l), & (1 \leq l \leq m), \\ 0, & (otherwise). \end{cases}$$

In this section, the detailed parameter settings are presented in the captions of each graph. We compared the proposed model (referred to as the variable batch service model in the following section to clarify the distinction) with the regular (i.e., fixed) batch size model [26] of size K. If $K = 1$, this becomes equivalent to Naor's model [23]. As mentioned in Section 3, the variable batch service model encompasses Naor's and the regular batch service models.

Notably, if no customers join the system according to the parameter settings, we assume that TH, SW, and MR become 0. First, we note the following remark from Lemma 6:

Remark 1. *$l^*(i)$ becomes a non-decreasing function (except for the case of $l^*(i) = 0$; that is, the customer balks in the setting of fees, that is, $p_l = f + s/l$.*

Table 1 summarizes the numerical results for the best batch size conditional on i, that is, the number of complete batches in the system when a tagged first-arriving customer arrives. Notably, 0 indicates that the best option for the customer is to balk. Clearly, the best batch size increases as i increases, as mentioned in Remark 1. This result leads us to presume that to prevent more serious traffic congestion, it is individually optimal for customers to tend to create large batches while sharing fees with other customers if the system is crowded.

Table 1. Best batch sizes conditional on i under $\lambda = 5, \mu = 2, R = 3, s = 1, s \in \{0.1, 1, 2\}$, $f \in \{0.1, 0.5\}, m = 5$ and $C = 1$.

(s, f)	$i = 0$	$i = 1$	$i = 2$	$i = 3$	$i = 4$
$(0.1, 0.1)$	1	1	2	3	0
$(0.1, 0.5)$	1	1	2	3	0
$(1.0, 0.1)$	2	3	4	5	0
$(1.0, 0.5)$	2	3	4	5	0
$(2.0, 0.1)$	3	4	5	5	0
$(2.0, 0.5)$	3	4	5	5	0

In addition, the distribution of B_{new}, that is $f_{\text{new}}(n)$, which is given by Theorem 3, is shown in Figure 5. We depict examples of $s \in \{0.1, 1, 2\}$ and $f = 0.1$. The other parameter settings are the same as those in the example in Table 1. These results can be utilized for real-world management of this system. For example, the probabilities that $B_{\text{new}} = 4, 5$ are 0 when $s = 0.1$ and $f = 0.1$. Therefore, three is sufficient for the capacity of cars, although the current setting for the maximum capacity of cars is $m = 5$. This finding enables the system administrator to reduce excessive costs.

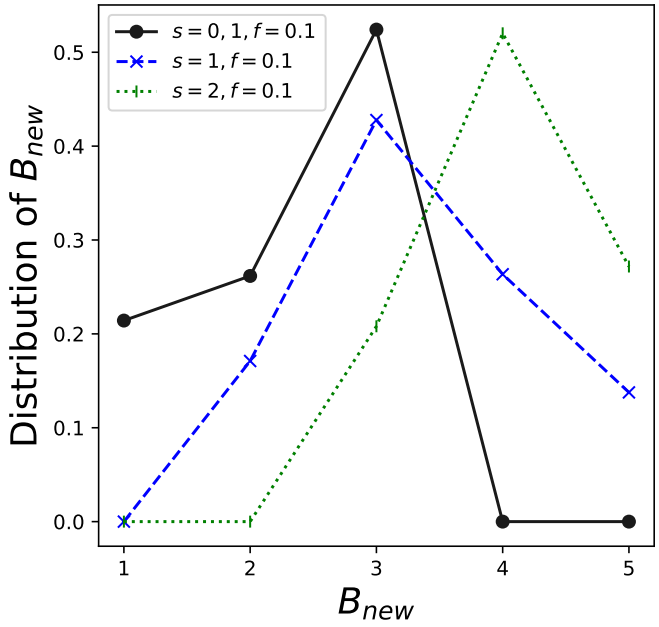
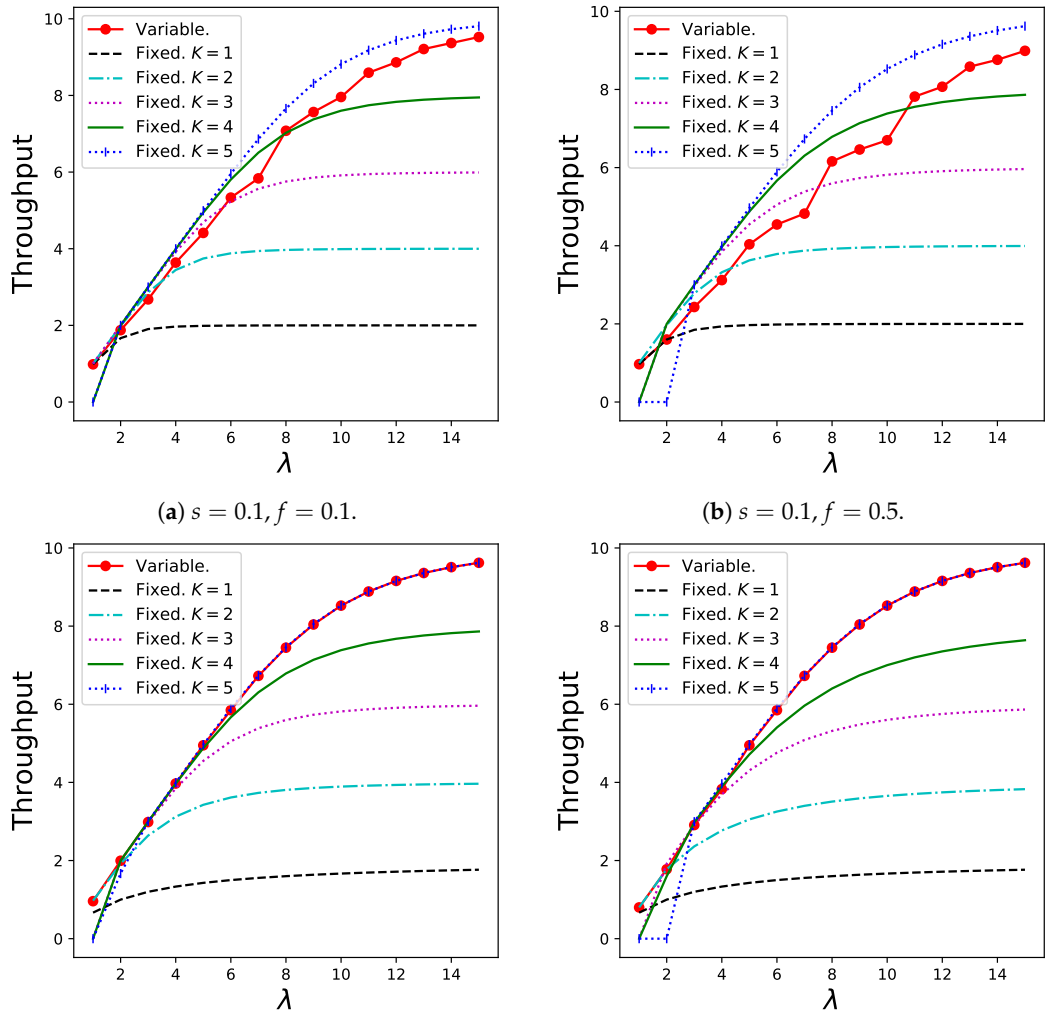


Figure 5. Distribution of B_{new} under $\lambda = 5, \mu = 2, R = 3, s = 1, s \in \{0.1, 1, 2\}, f = 0.1$, $m = 5$ and $C = 1$.

The results for the throughput TH in Theorem 6 are shown in Figure 6. Generally, the variable batch service model shows the best results for any arrival rate λ unless s is not extremely small because customers can choose the ideal batch sizes depending on λ, which prevents excessively long waiting times or traffic congestion that results in customers joining the system. Thus, the variable model is more robust to changing customer demands than the regular batch service model.

The reason for the poor performance of the variable batch service model when s is relatively low (Figure 6a,b) can be explained as follows: Customers tend to create small batches because the sharing fee that can be saved by creating a larger batch is less compared to the long expected waiting time in this case. This trend is presented in Table 1. However, if a considerably large number of small batches are generated, the congestion in the queue becomes serious, resulting in the balking of many customers. Therefore, a relatively large K (for example, $K = 5$) is more desirable in these graphs from the throughput viewpoint.

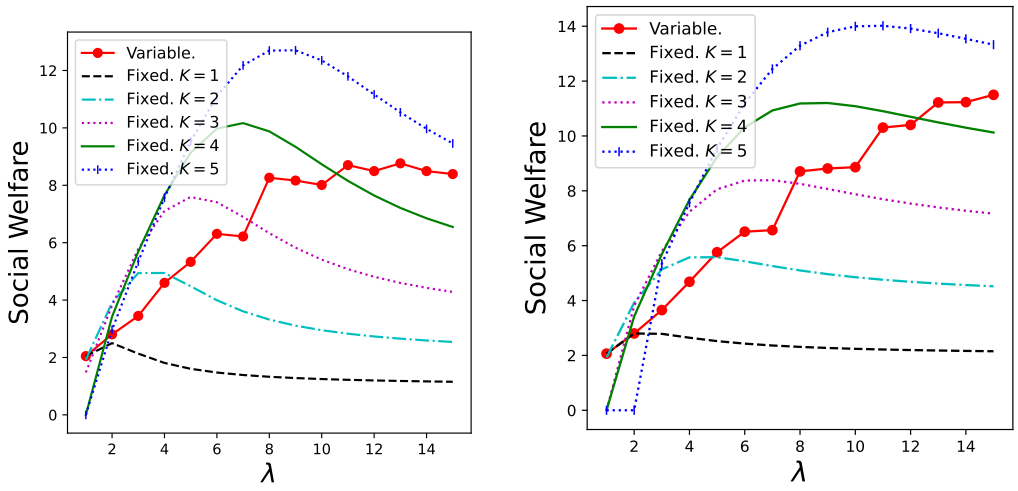
Figure 7 shows the results for social welfare, SW, in Theorem 6. The variable batch service model exhibited the best performance when s as well as TH are relatively high. When $s = 2$, the variable batch service model shows stable performance for any λ, whereas the regular batch service model with small batch sizes always shows poor performance, and the regular batch service model with large batch sizes shows poor results within the low λ zone. Therefore, the variable batch service model can be considered robust against demand fluctuations when s is large.



(**a**) $s = 0.1, f = 0.1$. (**b**) $s = 0.1, f = 0.5$.

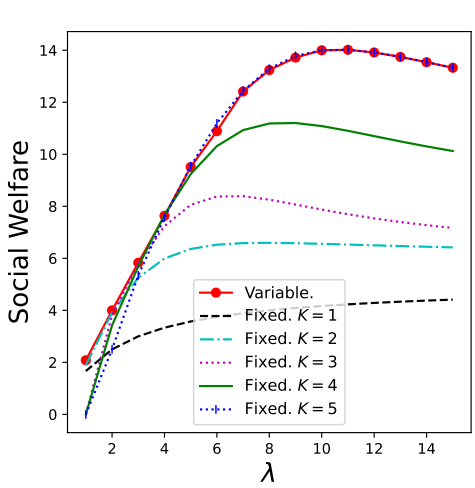
(**c**) $s = 2, f = 0.1$. (**d**) $s = 2, f = 0.5$.

Figure 6. Throughput for the variable batch service model under $\mu = 2, R = 3, s \in \{0.1, 2\}$, $f \in \{0.1, 0.5\}$ and $C = 1$, and that for the regular batch service model under $\mu = 2, R = 3$, $s \in \{0.1, 2\}, f \in \{0.1, 0.5\}, m = 5, C = 1$ and $K \in \{1, 2, 3, 4, 5\}$.

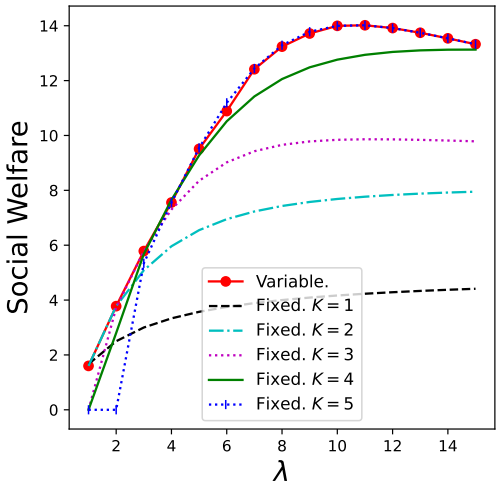


(**a**) $s = 0.1, f = 0.1$. (**b**) $s = 0.1, f = 0.5$.

Figure 7. *Cont.*



(**c**) $s = 2, f = 0.1$.

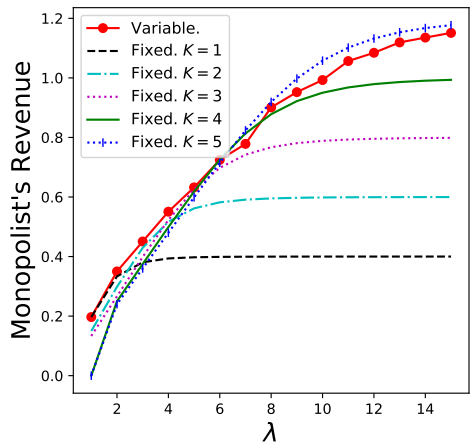


(**d**) $s = 2, f = 0.5$.

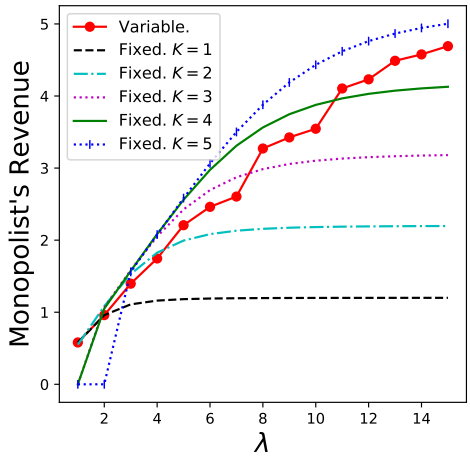
Figure 7. Social welfare for the variable batch service model under $\mu = 2, R = 3, s \in \{0.1, 2\}$, $f \in \{0.1, 0.5\}$ and $C = 1$, and that for the regular batch service model under $\mu = 2, R = 3$, $s \in \{0.1, 2\}, f \in \{0.1, 0.5\}, m = 5, C = 1$ and $K \in \{1, 2, 3, 4, 5\}$.

However, we note that the variable batch service model shows very poor performance when s is low, because batches are generated excessively and induce congestion in the system. In other words, a system where customers behave individually does not necessarily lead to a socially optimal state. Therefore, as a practical message, it is important for an administrator to set s to a relatively high value to avoid this situation.

The results for the monopolist's revenue, MR, are shown in Figure 8. The variable batch service model exhibits good performance as long as both s and f are relatively high (as in Figure 8d). However, especially when f is extremely small, the revenue of the variable batch service model decreases slightly.

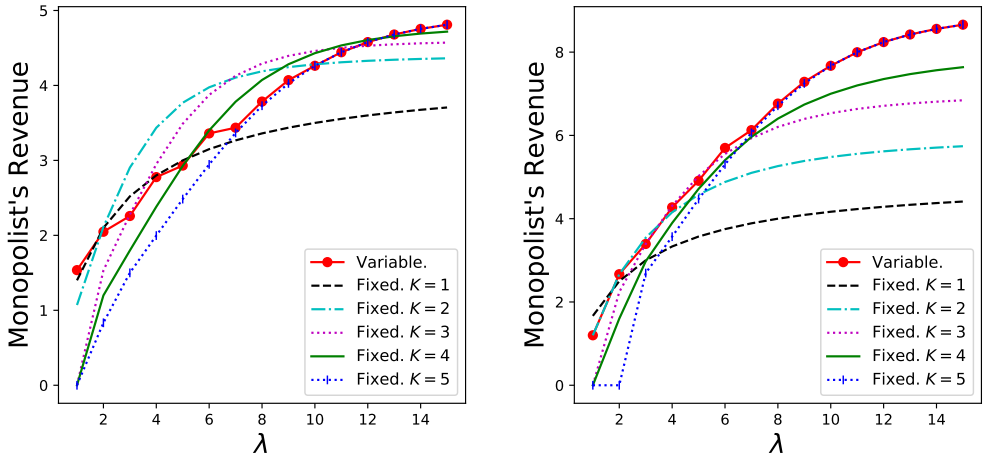


(**a**) $s = 0.1, f = 0.1$.



(**b**) $s = 0.1, f = 0.5$.

Figure 8. *Cont.*



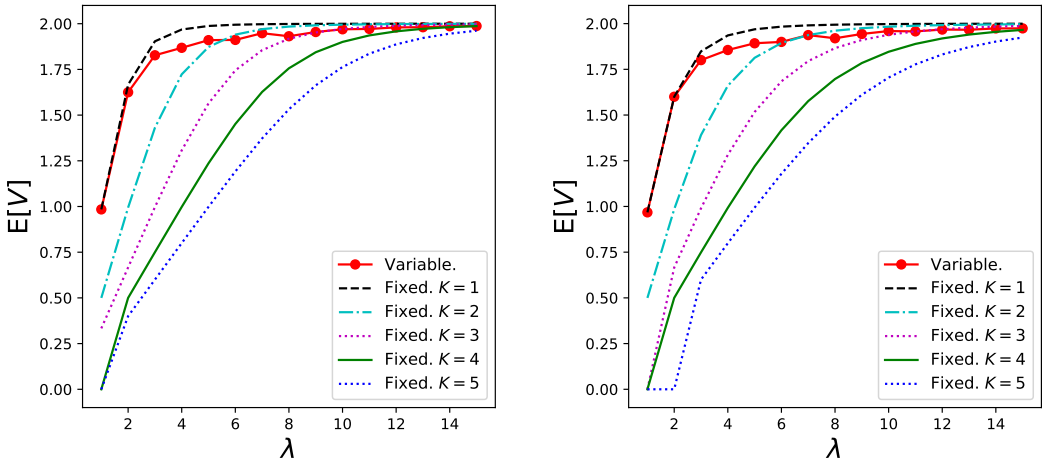
(**c**) $s = 2, f = 0.1$. (**d**) $s = 2, f = 0.5$.

Figure 8. Monopolist's revenue for the variable batch service model under $\mu = 2, R = 3$, $s \in \{0.1, 2\}, f \in \{0.1, 0.5\}$ and $C = 1$, and that for the regular batch service model under $\mu = 2, R = 3$, $s \in \{0.1, 2\}, f \in \{0.1, 0.5\}, m = 5, C = 1$ and $K \in \{1, 2, 3, 4, 5\}$.

The reasons for these results are as follows: In the variable batch-service model (compared with the regular batch service model), the following two effects were observed:

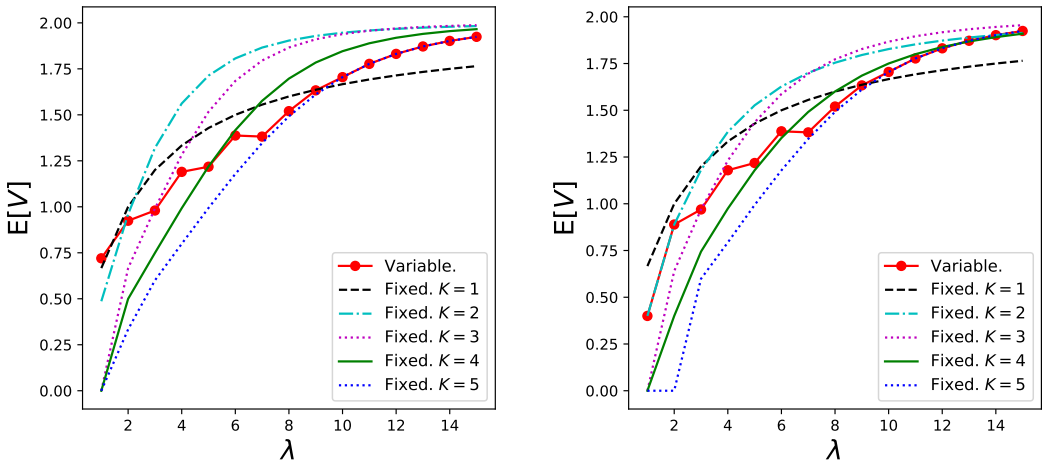
(i) The increase in the throughput.
(ii) The decrease in the number of batches in the system.

Effect (i) is presented in Table 1. Effect (ii) occurs because customers create larger batches if many batches already exist in the system to prevent serious congestion in the queue. When the fixed fee is low, the dominance of the total revenue is based only on the number of batches in the system that corresponds to the sharing fee. Specifically, the numerical results for the number of newly generated batches per unit time $\mathrm{E}[V]$ for the same parameters are shown in Figure 9. Therefore, in this case, effect (ii) increases, and revenue decreases. When a fixed fee is imposed to some extent, effect (i) increases as the arrival rate increases, and the total amount of fixed fees, which is proportional to the throughput, increases. Therefore, the administrator must set f as well as s to some high value.



(**a**) $s = 0.1, f = 0.1$. (**b**) $s = 0.1, f = 0.5$.

Figure 9. *Cont.*



(**c**) $s = 2, f = 0.1$. (**d**) $s = 2, f = 0.5$.

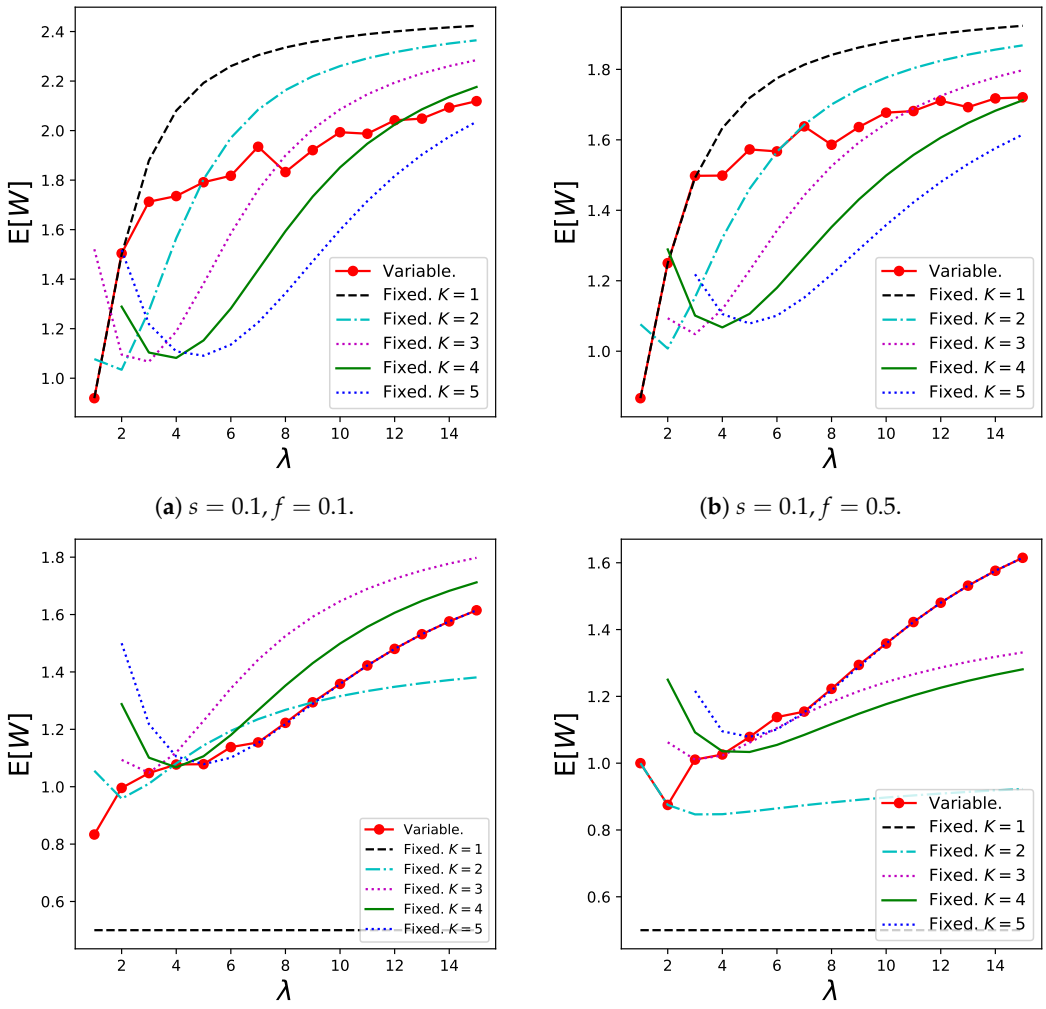
Figure 9. $\mathrm{E}[V]$ for the variable batch service model under $\mu = 2, R = 3, s \in \{0.1, 2\}$, $f \in \{0.1, 0.5\}$ and $C = 1$, and that for the regular batch service model under $\mu = 2, R = 3$, $s \in \{0.1, 2\}, f \in \{0.1, 0.5\}, m = 5, C = 1$ and $K \in \{1, 2, 3, 4, 5\}$.

Furthermore, it is worth noting that the variable batch service model is more suitable to optimize all of TH, SW, MR, simultaneously, compared to the regular batch service model, when the fees are properly set high ($s = 2$ and $f = 0.5$ in this example). See Figures 6d, 7d and 8d. For TH and SW, the variable batch service model and the regular batch service model with $K = 5$ show the best performances when $\lambda \geq 3$. For MR, the variable batch service model and the regular batch service model with $K = 5$ also show the best performances when $\lambda \geq 7$. On the other hand, when $4 \leq \lambda \leq 6$, the variable batch service model and the regular batch service model with $K = 3$ take better values in MR than that of the regular batch service model with $K = 5$. These results for the regular batch service model imply that the monopolist should set the batch size K smaller to maximize MR compared to the batch size which maximizes TH and SW. However, as has been shown, the variable batch service model enables all TH, SW, and MR to be the best values simultaneously as long as the fees are set high properly.

In addition, Figure 10 shows the numerical results for $\mathrm{E}[W]$, which are given by (16). The results for the variable batch service model generally increase as λ increases, whereas most results for the regular batch service model have minimum points. This is because customers can reduce their batch sizes to prevent extremely long waiting times when the arrival rate is low.

Notably, $\mathrm{E}[W]$ in the variable batch service model is lower than that in the regular batch service model in which TH and SW are close to those in the variable batch service model under the specific parameter setting, for example, when $\lambda = 3$ in Figures 6d, 7d and 10d, respectively. Therefore, from the waiting time perspective, a variable batch service model is preferred in this case as well. On the contrary, the regular batch service model with $K = 1$ shows the best performance when $\lambda = 1$ from the perspective of $\mathrm{E}[W]$ and the other three performance measures.

We note that the tendency of the results of $\mathrm{E}[W]$ does not necessarily coincide with that of SW in Figure 7. This is because $\mathrm{E}[W]$ does not reflect information of balking customers. For example, the results for $K = 1, 2$ within high λ zone in Figure 7d show extremely small values because the dominant customers abandon the service (as can be confirmed in Figure 6d). Therefore, a small $\mathrm{E}[W]$ does not always imply superiority from a system point of view. However, the results for $\mathrm{E}[W]$ are valuable from a customer's perspective. The proposed model is useful in obtaining numerous performance measures from various perspectives.



(**a**) $s = 0.1, f = 0.1$. (**b**) $s = 0.1, f = 0.5$.

(**c**) $s = 2, f = 0.1$. (**d**) $s = 2, f = 0.5$.

Figure 10. $\mathrm{E}[W]$ for the variable batch service model under $\mu = 2, R = 3, s \in \{0.1, 2\}$, $f \in \{0.1, 0.5\}$ and $C = 1$, and that for the regular batch service model under $\mu = 2, R = 3$, $s \in \{0.1, 2\}, f \in \{0.1, 0.5\}, m = 5, C = 1$ and $K \in \{1, 2, 3, 4, 5\}$.

5. Conclusions

In this study, we proposed a variable batch service model in which first-arriving customers can select the batch size flexibly to maximize their individual utility. We formulated this model as a three-dimensional Markov chain and created a book-type transition diagram. We then demonstrated several properties of the sojourn time of a tagged customer and the best batch sizes, conditional on the number of complete batches in the system. Based on these properties, we proposed an effective algorithm to construct a necessary and sufficient size of state space for the Markov chain provided that all customers adopt the threshold-type equilibrium strategy.

Furthermore, we proved that the best batch size, provided that a tagged customer observes i complete batches in the system upon arrival, is a non-decreasing function for i if the reward for the completion of a batch service with size l is an increasing function of l; in other words, the fee decreases as the batch becomes larger. This implies that to prevent more serious traffic congestion, creating large batches while sharing fees with other customers is individually optimal for customers if the system is crowded.

In the numerical experiment, we showed several results on performance measures, throughput, social welfare, monopolist's revenue, and expected sojourn time, and compared them with the regular batch service model [26]. The proposed model showed better

performance when the fixed and sharing fees were relatively high. In addition, the proposed model does not show extremely poor results for any arrival rate as long as both fees are set in a well-balanced manner compared to the regular batch size model. This indicates that the present model is a robust model for the fluctuation of customer demand. Moreover, we showed that the introduction of the variable batch service model can lead to optimizing all of the throughput, social welfare, and monopolist's revenue, simultaneously, compared to the regular batch service model when the fees were relatively high.

However, the proposed model performed poorly when fees were low. Therefore, a system in which customers select optimal batch sizes individually does not necessarily lead to a socially optimal state. Thus, it is important to set both fixed and sharing fees to some extent for the administrator in this system.

Further studies of the variable batch service model must consider the mechanism by which all customers (including non-first-arriving customers) can freely select their batch sizes. Analysis of this model is difficult; however, the change in performance is intriguing.

Author Contributions: Conceptualization, A.N. and T.P.-D.; methodology, A.N. and T.P.-D.; software, A.N.; validation, A.N. and T.P.-D.; investigation, A.N. and T.P.-D.; writing—original draft preparation, A.N.; writing—review and editing, T.P.-D.; supervision, T.P.-D.; project administration, T.P.-D.; funding acquisition, A.N. and T.P.-D. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by JSPS KAKENHI Nos. 21K11765, 18K18006, and 23KJ0249, JST SPRING No. JPMJSP2124. In addition, this study was also funded by F-MIRAI: R&D Center for Frontiers of MIRAI in Policy and Technology, the University of Tsukuba, and Toyota Motor Corporation collaborative R&D center.

Data Availability Statement: Not applicable.

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

References

1. Bailey, N.T. On queueing processes with bulk service. *J. R. Stat. Soc. Ser. B* **1954**, *16*, 80–87. [CrossRef]
2. Sasikala, S.; Indhira, K. Bulk service queueing models-a survey. *Int. J. Pure Appl. Math.* **2016**, *106*, 43–56.
3. Chaudhry, M.; Templeton, J.G. *First Course in Bulk Queues*; Wiley: Hoboken, NJ, USA, 1983.
4. Dshalalow, J.H. On modulated random measures. *J. Appl. Math. Stoch. Anal.* **1991**, *4*, 305–312. [CrossRef]
5. Dshalalow, J.H. A single-server queue with random accumulation level. *J. Appl. Math. Stoch. Anal.* **1991**, *4*, 203–210. [CrossRef]
6. Neuts, M.F. A general class of bulk queues with Poisson input. *Ann. Math. Stat.* **1967**, *38*, 759–770. [CrossRef]
7. Medhi, J. Waiting time distribution in a Poisson queue with a general bulk service rule. *Manag. Sci.* **1975**, *21*, 777–782. [CrossRef]
8. Borthakur, A. A Poisson queue with a general bulk service rule. *J. Assam Sci. Soc.* **1971**, *14*, 162–167.
9. Easton, G.; Chaudhry, M. The queueing system $E_k/M(a,b)/1$ and its numerical analysis. *Comput. Oper. Res.* **1982**, *9*, 197–205. [CrossRef]
10. Medhi, J. *Recent Developments in Bulk Queueing Models*; Wiley Eastern: Mumbai, India, 1984.
11. Powell, W.B. Stochastic Delays in Transportation Terminals: New Results in the Theory and Application of Bulk Queues. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, USA, 1981.
12. Chaudhry, M.L.; Madill, B.; Briere, G. Computational analysis of steady-state probabilities of $M/G(a,b)/1$ and related nonbulk queues. *Queueing Syst.* **1987**, *2*, 93–114. [CrossRef]
13. Pradhan, S.; Gupta, U. Analysis of an infinite-buffer batch-size-dependent service queue with Markovian arrival process. *Ann. Oper. Res.* **2019**, *277*, 161–196. [CrossRef]
14. Cosmetatos, G.P. Closed-form equilibrium results for the $M/M(a,\infty)/N$ queue. *Eur. J. Oper. Res.* **1983**, *12*, 203–204. [CrossRef]
15. Sim, S.; Templeton, J. Computational procedures for steady-state characteristics of unscheduled multi-carrier shuttle systems. *Eur. J. Oper. Res.* **1983**, *12*, 190–202. [CrossRef]
16. Sim, S.; Templeton, J.G.C. Further results for the $M/M(a,\infty)/N$ batch-service system. *Queueing Syst.* **1990**, *6*, 277–286. [CrossRef]
17. Chaudhry, M.L.; Gupta, U.C. Modelling and analysis of $M/G(a,b)/1/N$ queue–a simple alternative approach. *Queueing Syst.* **1999**, *31*, 95–100. [CrossRef]
18. Banerjee, A.; Gupta, U.C.; Chakravarthy, S.R. Analysis of a finite-buffer bulk-service queue under Markovian arrival process with batch-size-dependent service. *Comput. Oper. Res.* **2015**, *60*, 138–149. [CrossRef]
19. Chaudhry, M.; Abhijit, D.B.; Sitaram, B.; Veena, G. A novel computational procedure for the waiting-time distribution (in the queue) for bulk-service finite-buffer queues with poisson input. *Mathematics* **2023**, *11*, 1142. [CrossRef]

20. Briere, G.; Chaudhry, M. Computational analysis of single-server bulk-service queues, $M/G^Y/1$. *Adv. Appl. Probab.* **1989**, *21*, 207–225.
21. Nakamura, A.; Phung-Duc, T. A moment approach for a conditional central limit theorem of infinite-server queue: A case of $M/M^X/\infty$ queue. *Mathematics* **2023**, *11*, 2088. [CrossRef]
22. Pradhan, S.; Gupta, U.; Samanta, S. Queue-length distribution of a batch service queue with random capacity and batch size dependent service: $M/G_r^Y/1$. *Opsearch* **2016**, *53*, 329–343. [CrossRef]
23. Naor, P. The regulation of queue size by levying tolls. *Econom. J. Econom. Soc.* **1969**, *37*, 15–24. [CrossRef]
24. Hassin, R.; Haviv, M. *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*; Springer Science & Business Media: New York, NY, USA, 2003; Volume 59.
25. Hassin, R. *Rational Queueing*; CRC Press: Boca Raton, FL, USA, 2016.
26. Bountali, O.; Economou, A. Equilibrium joining strategies in batch service queueing systems. *Eur. J. Oper. Res.* **2017**, *260*, 1142–1151. [CrossRef]
27. Bountali, O.; Economou, A. Equilibrium threshold joining strategies in partially observable batch service queueing systems. *Ann. Oper. Res.* **2019**, *277*, 231–253. [CrossRef]
28. Bountali, O.; Economou, A. Strategic customer behavior in a two-stage batch processing system. *Queueing Syst.* **2019**, *93*, 3–29. [CrossRef]
29. Nakamura, A.; Phung-Duc, T. Strategic customer behaviors in observable multi-server batch service queueing systems with shared fee and server maintenance cost. In Proceedings of the Performance Evaluation Methodologies and Tools, Online, 16–18 November 2022; Springer: Cham, Switzerland, 2023; pp. 3–13.
30. Wang, Z.; Liu, L.; Shao, Y.; Chai, X.; Chang, B. Equilibrium joining strategy in a batch transfer queuing system with gated policy. *Methodol. Comput. Appl. Probab.* **2020**, *22*, 75–99. [CrossRef]
31. Manou, A.; Economou, A. Equilibrium balking strategies for a clearing queueing system in alternating environment. *Ann. Oper. Res.* **2013**, *208*, 489–514.
32. Canbolat, P.G. Bounded rationality in clearing service systems. *Eur. J. Oper. Res.* **2020**, *282*, 614–626. [CrossRef]
33. Manou, A.; Economou, A.; Karaesmen, F. Strategic customers in a transportation station: When is it optimal to wait? *Oper. Res.* **2014**, *62*, 910–925. [CrossRef]
34. Manou, A.; Canbolat, P.G.; Karaesmen, F. Pricing in a transportation station with strategic customers. *Prod. Oper. Manag.* **2017**, *26*, 1632–1645. [CrossRef]
35. Logothetis, D.; Economou, A. The impact of information on transportation systems with strategic customers. *Prod. Oper. Manag.* **2023**, *32*, 2189–2206. [CrossRef]
36. Czerny, A.I.; Guo, P.; Hassin, R. Shall firms withhold exact waiting time information from their customers? A transport example. *Transp. Res. Part B Methodol.* **2022**, *166*, 128–142. [CrossRef]
37. Calvert, B. The Downs-Thomson effect in a Markov process. *Probab. Eng. Inf. Sci.* **1997**, *11*, 327–340. [CrossRef]
38. Afimeimounga, H.; Solomon, W.; Ziedins, I. The Downs-Thomson paradox: Existence, uniqueness and stability of user equilibria. *Queueing Syst.* **2005**, *49*, 321–334. [CrossRef]
39. Afimeimounga, H.; Solomon, W.; Ziedins, I. User equilibria for a parallel queueing system with state dependent routing. *Queueing Syst.* **2010**, *66*, 169–193. [CrossRef]
40. Chen, Y.; Holmes, M.; Ziedins, I. Monotonicity properties of user equilibrium policies for parallel batch systems. *Queueing Syst.* **2012**, *70*, 81–103. [CrossRef]
41. Wang, A.; Ziedins, I. Probabilistic selfish routing in parallel batch and single-server queues. *Queueing Syst.* **2018**, *88*, 389–407. [CrossRef]
42. Logothetis, D.; Economou, A. Routing of strategic passengers in a transportation station. In Proceedings of the Performance Engineering and Stochastic Modeling, Online, 9–10 and 13–14 December 2021; Springer: Cham, Switzerland, 2021; pp. 308–324.
43. Nguyen, Q.H.; 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. [CrossRef]
44. Afeche, P.; Mendelson, H. Pricing and priority auctions in queueing systems with a generalized delay cost structure. *Manag. Sci.* **2004**, *50*, 869–882. [CrossRef]
45. Li, Q.; Guo, P.; Wang, Y. Equilibrium analysis of unobservable $M/M/n$ priority queues with balking and homogeneous customers. *Oper. Res. Lett.* **2020**, *48*, 674–681. [CrossRef]
46. Van Woensel, T.; Vandaele, N. Modeling traffic flows with queueing models: A review. *Asia-Pac. J. Oper. Res.* **2007**, *24*, 435–461. [CrossRef]