Mathematics

http://www.aimspress.com/journal/Math

AIMS Mathematics, 8(11): 25382–25408.
DOI:10.3934/math.20231295
Received: 05 June 2023
Revised: 16 August 2023
Accepted: 24 August 2023
Published: 31 August 2023

Research article

Improving efficiency of the queueing system with two types of customers by service decomposition

Linhong Li[1], **Wei Xu**[1], **Zhen Wang**[2] **and Liwei Liu**[1,*]

[1] School of Mathematics and Statistics, Nanjing University of Science and Technology, Nanjing, Jiangsu 210094, China

[2] School of Mathematics and Computing Science, Guilin University of Electronic Technology, Guilin 541004, China

* **Correspondence:** Email: NJUSTliuliwei@163.com.

Abstract: The economic improvements of a queueing system with two types of customers achieved by service decomposition are examined. The service process for a Type 2 customer can be split into two phases: a basic service and an additional service. The basic service rate is equal to that of the Type 1 customer. Additional services can be viewed as orders stored in inventory and processed when the server is idle. Once a new customer arrives during idle time, the server will provide a basic service to the customer upon his/her arrival. That is, customers have preemptive priority for orders stored in inventory. We obtain a two-dimensional unbounded Markov process and apply the multivariate generating function to derive the stationary probability of the proposed model as well as some performance measures. The condition under which performing service decomposition can improve economic efficiency is also obtained. Both the optimal inventory capacity and the minimum system cost are obtained numerically. Numerical experiments demonstrate the impact of optimal inventory setting on economic improvement efficiency. Finally, simulation experiments prove the correctness of our theoretical analysis.

Keywords: queueing-inventory system; service decomposition; economic improvement; two types of customers
Mathematics Subject Classification: 60K25, 68M20, 90B05

1. Introduction

Classical queueing-inventory systems often assume that when customer service is complete, the system inventory level decreases accordingly. When inventory is depleted, replenishment orders are then triggered. This replenishment is provided by an external supplier. We recommend Marand et al. [1] to readers interested in a comprehensive look at the queueing-inventory system. Furthermore, the service system can also produce the products by itself. We want to highlight the preliminary service (PS) that is performed by servers during idle periods to generate stockable inventory items and consumed as the customer's service is completed. PS is first introduced in Hanukov et al. [2], which is executed when the server is empty for the customer. The authors gave several cases where executing PS during the server's idle time can increase the idle time fraction of the system. Hanukov et al. [3] extended the research of Hanukov et al. [2] and analyzed the performance improvement of the PS on the service system in which the inventory produced by PS deteriorates while in storage, creating spoilage costs. The other one is a complementary service that can be conducted without customers. Complementary service is presented for the first time in Hanukov [4]. The service is divided into two phases: the opening service and the complementary service. The customer's complementary service is stored in the inventory as a pending order that is processed when the server is dormant. They compared the proposed model with the status quo model under different cost scenarios regarding efficiency improvement. The above two services can reduce the overall sojourn time of the customer by performing a part of the service when the server is dormant.

Malachowski and Simonini [5] pointed out that time wasted by employees at work was still costing companies billions, which further emphasized the importance of improving the efficiency of service systems. Using the system' idle time to perform other services is prevalent for service systems to improve efficiency. Many authors proposed "vacation" models to address the service system issue. When the system is empty, the server does not remain dormant but handles service-related or other auxiliary tasks (e.g., maintenance, repair, or organization). Doshi [6] gave some methods and decomposition results of the queueing systems with vacation. Ke [7] investigated an M/G/1 queueing system in which the authors analyzed the optimal setting of thresholds for the server to take vacations. Zhang et al. [8] studied a queueing-inventory system with server vacations. The authors assumed that the server goes for multiple vacations once the system inventory is depleted. Due to vacations, the servers are not immediately available to serve customers who arrive during vacations. To reduce the impact of vacations on the system's primary services, many authors have addressed this by limiting the number of vacations (Meena et al. [9] and Ke [10]) or introducing working vacations to queueing systems (Laxmi et al. [11] and Tian [12]). The above articles all assumed that the tasks performed by the server during the vacation periods are not related to the primary task. Hanukov et al. [2–4] also studied the utilization of the idle time to improve the efficiency of service systems. In contrast, the work done by the server during idle periods is related to the primary task and can reduce the customer's overall sojourn time.

In this article, we study queueing systems with two types of customers. There are already well-established theoretical results for queueing systems with multiple classes of customers. Blanc et al. [13] described an M/M/c queueing system with two arrival streams. One type of customer can be rejected for entry. They also obtained a stationary admission policy to maximize the discounted cost function. Turhan et al. [14] also studied the optimal admission policy for a queueing system with two types of customers. When the system is full, the arrival of a Type 1 customer terminates the service of a Type 2 customer. The authors gave the admission policy for Type 2 customers in a threshold style. Kim and Kim [15] investigated a retrial queueing system with two classes of customers in which class-2 customers enter the retrial orbit when the system is busy, and class-1 customers form an infinite queue in the system. Both the waiting time distributions of class-1 and class-2 customers are obtained. We place particular emphasis on Hanukov [16]. The author studied a queueing-inventory model with two types of customers: skeptical and trusting customers. In this article, the trusting customer's service is divided into two phases: opening service and complementary service. A complementary service is stored in the system to be processed when the server is idle. Customers who insist on receiving the complete service are defined as skeptical customers. The system stores a finite number of complementary services. Once the system is fully stocked, both types of customers receive full services with the same service rate. The author compared the proposed model with a classical M/M/1 model regarding economic efficiency.

In this work, we investigate improving the efficiency of a service system with two types of customers by service decomposition. Hanukov [16] also studied a queueing system with two types of customers and finite inventory capacity. The author assumed that skeptical customers refuse to decompose their services when the inventory is not full. Once the inventory is full, all customers are considered skeptical and served at the same rate. In this article, we also investigate a queueing model with limited inventory capacity and two types of customers, Model 3. We assume that the complete service rate differs for the two customer types. Before the system inventory is full, all customers are served at the same rate. After a Type 2 customer leaves, the system inventory increases by one. Once the inventory is full, the service process of Type 2 customers is no longer decomposed. Both types of customers leave the system after receiving the complete service with different service rates. The whole model becomes a queueing-inventory model. Our model is mostly motivated by the following examples. Customers choose the necessary products in a furniture company and settle at the cashier. Customers who only purchase small items leave the system after checkout (Type 1 customers); customers who purchase large items need the merchant to provide delivery service after checkout (Type 2 customers). The service of Type 2 customers can be divided into two phases: one is the checkout process and the other is delivery. Another example of service decomposition is ultrasound in hospitals. The whole ultrasound process includes the examination and the report's writing afterward. In some primary hospitals, the doctor writes a report immediately after examining the patient. The patient can leave the hospital at once with the report. In some tertiary hospitals, however, the doctor will finish the examination for all waiting patients and then write the report at leisure.

The major scientific contributions made in this paper can be summarized in the following:

(1) We study a queueing system with two types of customers. In this case, the service for Type 2 customers can be decomposed into two phases. We obtain closed-form expressions for the stationary distribution of the queueing system with service decomposition applied and explicit expressions to the system performance measures.

(2) Model 1 is a two-unbounded queueing system, which is rare in the literature. Combing the probability generating function method with the multivariable *L′Hôpital* rule, we derive the explicit expression of the mean number of stored orders in the system. This approach has been introduced for the first time in [4].

(3) The performance measures for the queueing model without service decomposition are derived by the probability generating function. After constructing an appropriate cost function, we obtain the condition for determining the adoption of the service decomposition. Under this condition, service decomposition should be provided for all Type 2 customers. In addition, the fraction of system idle time can become larger under a specific condition by service decomposition.

(4) We propose a queueing system with service decomposition and finite inventory capacity. Taking the inventory capacity as the independent variable and using an optimization algorithm, we can obtain the optimal inventory capacities and the minimum system cost under different parameter settings. Moreover, we also show the effect of each parameter on the optimal values through numerical examples. The economic improvements of the queueing system with service decomposition and optimal capacity are also analyzed numerically.

The rest of this paper is organized as follows. In Section 2, we give a brief description of the proposed model and obtain the closed-form expressions for the stationary probabilities and some system performance measures. In Section 3, we obtain the performance measures of the original system. An analysis of model selection in extreme cases is obtained by employing the constructed cost function. Section 4 includes a steady-state analysis of the limited capacity queueing model and an economic comparison with the previous two models. In order to validate our analysis results, comparisons of simulation and theoretical results are presented in Section 5. Finally, a summary is given in Section 6.

2. Model description

In this section, we thoroughly examine the Model 1 obtained by performing the service decomposition for all Type 2 customers. Furthermore, we obtain the system performance measures explicitly.

We consider an M/M/1 queueing system with two types of customers. The customers' arrival process is a Poisson process with rate λ. The service time of Type 1 customers is exponentially distributed with mean $1/\alpha$. The full service time of Type 2 customers follows an exponential distribution with mean $1/\mu$. In the previous literature, customer service is continuous. Even with staged service, customers are served on a global first-come, first-served order. However, we split the service of Type 2 customers into two phases: basic service and additional service. The basic service rate is equal to the service rate of Type 1 customers α. Additional services are stored in inventory as pending orders, waiting to be processed when the server is idle. After performing service decomposition for all Type 2 customers, a Type 2 customer receives a basic server and leave the system (so effectively acting as a Type 1 customers). The difference is that the departure of a Type 1 customer does not result in an increase in the number of additional services stored in inventory. Therefore, all customers can be seen as leaving the system immediately after receiving the basic service. Additionally, we assume that the proportion of Type 2 customers among all customers is q. Thus, after the customer's basic service is completed, the system inventory is increased by 1 with probability q. The additional service time is assumed to be exponentially distributed with rate β. We do not assume that the total mean duration of a decomposed service $1/\alpha + 1/\beta$ equals to the average full service time for a Type 2 customer $1/\mu$. See Section 3.1.2. The stored orders are executed only when no customers are in the system. Otherwise the server delays additional services and holds the orders in the inventory. In addition, when a customer arrives at a busy server processing the additional services, the undergoing service is interrupted immediately and the newly arriving customer receives service instead. In other words, customers have preemptive priority over stored orders. Therefore, the queueing behavior of customers in the system is consistent with the conventional M/M/1 queueing model with the traffic intensity λ/α.

Denote the number of customers and the stored orders in the system at time t by $N^1(t)$ and $I^1(t)$, respectively. Then, $X^1(t) = \{(N^1(t), I^1(t)), t \geq 0\}$ is a two-dimensional continuous-time Markov process with the state space $\Omega^1 = \{(n, i), n, i \in \mathbb{N}\}$. The transition situations between the states are shown in Figure 1. When the traffic intensity $\rho = \frac{\lambda}{\alpha} + \frac{\lambda q}{\beta} = \frac{\lambda\beta+\lambda\alpha q}{\alpha\beta} < 1$, the two-unbonded Markov process $\{X^1(t), t \geq 0\}$ is irreducible and recurrent with invariant probability $p^1_{n,i}, (n, i) \in \Omega^1$.

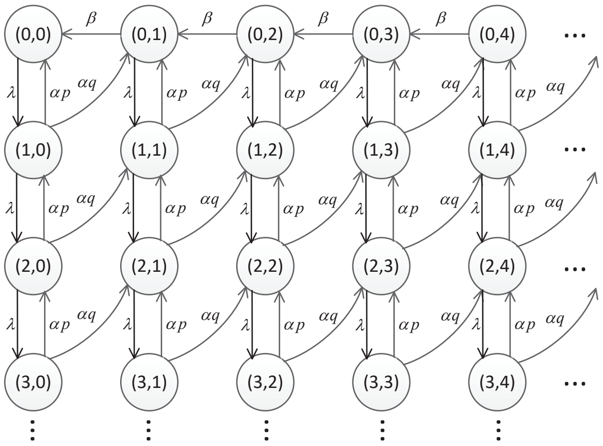


Figure 1. The state-transition diagram of the queueing inventory system.

2.1. Stationary analysis

The steady-state probability $p^1_{n,i}$ is defined as

$$p^1_{n,i} = \lim_{t\to\infty} Pr(N^1(t) = n, I^1(t) = i), (n, i) \in \Omega^1. \tag{2.1}$$

The balance equations between the states are as follows:

$$\lambda p^1_{0,0} = \beta p^1_{0,1} + p\alpha p^1_{1,0}, \tag{2.2}$$

$$(\alpha + \lambda)p^1_{n,0} = p\alpha p^1_{n+1,0} + \lambda p^1_{n-1,0},\ n \geq 1, \tag{2.3}$$

$$(\lambda + \beta)p^1_{0,i} = \beta p^1_{0,i+1} + p\alpha p^1_{1,i} + q\alpha p^1_{1,i-1},\ i \geq 1, \tag{2.4}$$

$$(\lambda + \alpha)p^1_{n,i} = \lambda p^1_{n-1,i} + p\alpha p^1_{n+1,i} + q\alpha p^1_{n+1,i-1},\ n \geq 1,\ i \geq 1. \tag{2.5}$$

The probability generating function is defined as $H(z, w) = \sum_{n=0}^{\infty}\sum_{i=0}^{\infty} p^1_{n,i} z^n w^i$, which is a multivariate function. The corresponding partial probability generating function is defined as $H_i(z) = \sum_{n=0}^{\infty} p^1_{n,i} z^n$, $i \geq 0$. We can obtain a set of equations in the following by multiplying the Eqs (2.2)–(2.5) by z^n and summing over all n.

$$[(\lambda + \alpha - \lambda z)z - p\alpha]H_0(z) = (\alpha z - p\alpha)p^1_{0,0} + \beta z p^1_{0,1}, \tag{2.6}$$

$$[(\alpha + \lambda - \lambda z)z - p\alpha]H_i(z) - q\alpha H_{i-1}(z) = (\alpha z - \beta z - p\alpha)p^1_{0,i} - q\alpha p^1_{0,i-1} + \beta z p^1_{0,i+1},\ i \geq 1. \tag{2.7}$$

We also define a generating function of boundary probabilities $p^1_{0,i}$, $i \geq 0$ as $B_0(w) = \sum_{i=0}^{\infty} p^1_{0,i} w^i$. Same as before, we multiply the Eq (2.6) by w^0 and Eq (2.7) by w^i. Then, after arranging, we get the probability generating function has the following form:

$$H(z, w) = \frac{[(\alpha z - \beta z - p\alpha - q\alpha w)w + \beta z]B_0(w) + \beta z(w-1)p^1_{0,0}}{w[(\alpha + \lambda(1-z))z - p\alpha - q\alpha w]}. \tag{2.8}$$

Note that there are two unknowns in Eq (2.8): $B_0(w)$ and $p^1_{0,0}$. In order to calculate $p^1_{0,0}$, we should first point out that when the system is stable, the inflow and outflow between the two state sets are balanced. Then, we focus on the transitions between rows in Figure 1. The balance equations between rows can be summarized as follows:

$$\lambda p^1_{n,\cdot} = q\alpha p^1_{n+1,\cdot} + p\alpha p^1_{n+1,\cdot} = \alpha p^1_{n+1,\cdot},\ n = 0, 1, 2, \cdots, \tag{2.9}$$

where $p^1_{n,\cdot} = \sum_{i=0}^{\infty} p^1_{n,i}$. Summing the Eq (2.9) over all n, we obtain $p^1_{0,\cdot} = 1 - \frac{\lambda}{\alpha}$. Then, we focus on the transitions between volumes. Same as the previous operation, the balance equations between volumes can be summarized as follows:

$$q\alpha(p^1_{\cdot,i} - p^1_{0,i}) = \beta p^1_{0,i+1},\ i = 0, 1, 2, \cdots, \tag{2.10}$$

Summing the Eq (2.10) over all i, we obtain

$$p^1_{0,0} = \frac{(\alpha - \lambda)\beta - q\alpha\lambda}{\alpha\beta}. \tag{2.11}$$

It should be reminded that we cannot get the explicit expression of $B_0(w)$. However, we will prove that any order partial derivatives of $B_0(w)$ can be derived in a particular procedure which will be shown later. Although $B_0(w)$ is an unknown function, we still give the expression for $H(z, w)$ as follows by substituting Eq (2.11) into Eq (2.8):

$$H(z, w) = \frac{[(\alpha z - \beta z - p\alpha - q\alpha w)w + \beta z]\alpha B_0(w) + z(w-1)((\alpha - \lambda)\beta - \lambda q\alpha)}{\alpha w[(\alpha + \lambda(1-z))z - p\alpha - q\alpha w]}. \tag{2.12}$$

Under the stability condition, we can derive several performance measures in following.

Theorem 2.1. *The mean number of customers in the system $E(N_1)$ is*

$$E(N_1) = \frac{\lambda}{\alpha - \lambda}. \tag{2.13}$$

Proof. The probability generating function is defined as $H(z, w) = \sum_{n=0}^{\infty} \sum_{i=0}^{\infty} p^1_{n,i} z^n w^i$. The mean number of customers in the system can be calculated by $\frac{\partial H(z,w)}{\partial z}|_{z=1,w=1}$. It leads to $\frac{0}{0}$. We apply the multivariable *L'Hôpital* rule developed by Lawlor [17] to it at the fixed direction $v = (1, 0)$. The result is obvious. □

Remark 1. Actually, as the server stop processing the stored orders when the customers arrive at the system, the queueing behavior of customers in this system is no different from the conventional M/M/1 queueing system with service intensity λ/α. Therefore, $E(N_1)$ is equal to the mean queue length of conventional M/M/1 queueing system with traffic intensity $\frac{\lambda}{\alpha}$.

Theorem 2.2. *The mean number of stored orders in the system $E(I)$ is*

$$E(I_1) = \frac{\partial H(z,w)}{\partial w}\Big|_{z=1,w=1} = \frac{\lambda q(\lambda\beta + \alpha^2 - \lambda\alpha p)}{(\alpha - \lambda)(\alpha\beta - \lambda(\beta + q\alpha))}. \tag{2.14}$$

Proof. According to the properties of multivariate probability generating function, the mean number of stored orders in the system $E(I_1) = \frac{\partial H(z,w)}{\partial w}\Big|_{z=1,w=1}$. Obviously, the partial derivative of the numerator with respect to w needs to use the expression of $B_0^{'}(w)$. Since $B_0(w)$ is not an explicit expression, we first need to obtain the expression of $B_0^{'}(w)$.

As mentioned before, we use the multivariable *L′Hôpital* rule to calculate $E(N_1)$ by taking derivations with respect to a fixed direction $\overrightarrow{\boldsymbol{v_1}} = (1,0)$. We can also apply the multivariable *L′Hôpital* to obtain $E(N_1)$ via a different direction $\overrightarrow{\boldsymbol{v_2}} = (0,1)$. The following symbols are used to simplify the expression: $H(z,w) = \frac{B+C}{A}$, where

$$A = \alpha w[(\alpha + \lambda(1 - z))z - p\alpha - q\alpha w], \tag{2.15}$$

$$B = [(\alpha z - \beta z - p\alpha - q\alpha w)w + \beta z]\alpha B_0(w), \tag{2.16}$$

$$C = z(w - 1)((\alpha - \lambda)\beta - \lambda q\alpha). \tag{2.17}$$

Then, we apply the multivariable *L′Hôpital* rule twice to $\frac{\frac{\partial(B+C)}{\partial z}A - \frac{\partial A}{\partial z}(B+C)}{A^2}$ via the fixed direction $\overrightarrow{\boldsymbol{v_2}}$=(0, 1), which is the partial derivative of the above numerator and denominator with respect to w. After substituting $z = 1$, $w = 1$ into the result, the mean number of customers in the system $E(N_1)$ also has the following expression:

$$E(N_1) = \frac{(\alpha\beta - \lambda\beta - \lambda\alpha q)B_0^{'}(w = 1) + \lambda q(\lambda - \alpha p)}{q^2\alpha^2}. \tag{2.18}$$

By comparing the Eqs (2.13) and (2.18), we can derive that

$$B_0^{'}(w = 1) = \frac{\lambda q[(\alpha - \lambda)^2 + \lambda q\alpha]}{(\alpha - \lambda)[\alpha\beta - \lambda(\beta + q\alpha)]}. \tag{2.19}$$

Then, $E(I_1) = \frac{\partial H(z,w)}{\partial w}\Big|_{z=1,w=1}$ also leads to $\frac{0}{0}$. Applying the multivariable *L′Hôpital* rule twice to the equation in direction $\overrightarrow{\boldsymbol{v_2}} = (0,1)$, the mean number of stored orders in the system $E(I_1)$ is given by

$$E(I_1) = \frac{(\beta + q\alpha)B_0^{'}(w = 1) - \lambda q}{q\alpha}. \tag{2.20}$$

Finally, by substituting the Eq (2.19) into Eq (2.20), the proof is concluded. □

Denote the mean sojourn time of a customer in the system by $E(S_1)$. Using the Little's law, we can obtain $E(S_1)$ as: $E(S_1) = E(N_1)/\lambda$. Similarly, we can obtain the mean sojourn time of a stored order in the system $E(S_o) = E(I_1)/\alpha_e$, where α_e represents the actual arrival rate of the stored orders. Obviously, $\alpha_e = q\alpha(1 - p_{0,.}^1) = \lambda q$ is also the arrival rate of the Type 2 customers when the system is stable. As for the steady-state probabilities $p^1_{n,i}$, $(n,i) \in \Omega^1$, we can calculate them according to the properties of the multivariate probability generating function $H(z,w)$ as follows:

$$p^1_{n,i} = \frac{\partial^{n+i} H(z,w)}{\partial^n z \partial^i w}\bigg|_{z=0,w=0} \cdot (n!i!)^{-1}. \tag{2.21}$$

It should be noted that the values of derivatives of $B_0(w)$ at $w = 0$: $B_0^{(k)}(w = 0)$, $k = 0,1,2,\cdots,i$ must be known before calculating $p^1_{n,i}$. We have obtained $B_0(w = 0) = p^1_{0,0}$. In Theorem 2.2, we apply the multivariable *L'Hôpital* rule to calculate $B_0^{'}(w = 1)$. Similarly, we can obtain $B_0^{(k)}(w = 0)$, $k = 0,1,2,\cdots,i$ by applying the multivariable *L'Hôpital* rule. Although this method is cumbersome, it is feasible.

Now, we focus on the conditional expected sojourn time $T(k, j)$ for a stored order when it enters the inventory with k customers in the system and another $j-1$ orders queueing before it. We denote this order as a tagged order. Based on the transition diagram, we can obtain the following theorem.

Theorem 2.3. *If the system state is* $(k, j-1)$ *when this tagged order is created, its conditional expected sojourn time* $T(k, j)$ *can be obtained as*

$$T(k, j) = \frac{k}{\alpha-\lambda} + \frac{j\cdot\alpha}{\beta(\alpha-\lambda)},\ k \geq 0, j \geq 1. \tag{2.22}$$

Proof. Before proving this theorem, we first introduce the concept of k-order busy period M_k for a classical M/M/1 queueing system. M_k refers to the duration from when there are K customers in the system to when there are no customers in the system. From Cohen [18], the M_k has the following expression: $M_k = \frac{k}{\alpha-\lambda}$, $k = 1,2,3,\cdots$. The conditional sojourn time of an order, given that it sees the state $(0, j-1)$ when it enters the inventory, can be calculated in the following. For $j = 1$, this order will be processed at first when the server is idle. Since the customer's basic service has preemptive priority over the additional service of the orders stored in inventory, additional service is interrupted by the arrival of a customer. Customers arriving during this basic service time will also be served. The time period from an order is interrupted until the server resumes additional service for the order can be viewed as the 1-order busy period M_1 of the $M/M/1$ queueing system. We obtain the $T(0,1)$ as follows:

$$T(0,1) = \frac{1}{\lambda+\beta} + \frac{\lambda}{\lambda+\beta}(M_1 + T(0,1)) = \frac{1}{\beta} + \frac{\lambda}{\beta}M_1 = \frac{\alpha}{\beta(\alpha-\lambda)}.$$

If arrival occurs before service completion with probability $\frac{\lambda}{\lambda+\beta}$, the order must wait an entire busy period M_1 before being served again. If no customer arrives while the order is being processed, then the order is directly served without interruption.

For $j \geq 2$, the average sojourn time $T(0, j)$ is obtained as

$$T(0, j) = \frac{1}{\lambda+\beta} + \frac{\lambda}{\lambda+\beta}(M_1 + T(0, j)) + \frac{\beta}{\lambda+\beta}T(0, j-1) = j \cdot \frac{\alpha}{\beta(\alpha-\lambda)}. \tag{2.23}$$

When the tagged order is created, there are k customers present in the system. Then, these k customers, as well as new arrivals during the service period, will be served before the server can process the inventory orders. This period can be regarded as the $k-$order busy period M_k of the classical M/M/1 queueing model. We can use the same method to obtain $T(k, j),\ k \geq 1,\ j \geq 1$.

$$T(k, j) = M_k + T(0, j) = \frac{k}{\alpha - \lambda} + \frac{j \cdot \alpha}{\beta(\alpha - \lambda)}. \tag{2.24}$$

The proof is completed. □

Remark 2. The conditional expected sojourn time for an order can help Type 2 customers understand the status of the order. Type 2 customers are homogeneous and risk neutral. A risk neutral customer will maximize his/her revenue by considering the waiting costs of additional services. Excessive inventory orders will delay the completion of additional service for Type 2 customers. Therefore, the inventory capacity of the system is limited when Type 2 customers can choose whether to accept the service decomposition or not.

3. Model 2: the original queueing model without service decomposition

In order to determine whether performing service decomposition for Type 2 customers improves the economic efficiency of the system, we further investigate the original model without service decomposition. If we do not split the service of Type 2 customers into two phases, the system is simplified to a classical M/M/1 queueing system with two types of customers in which all customers join a single waiting queue. Finally, we will compare the cost functions of the two models and obtain the condition for performing the service decomposition. Unlike Model 1, the customers' service in this system is continuous. The service time of the Type 1 customer follows an exponential distribution with parameter α. The service time of the Type 2 customer is exponentially distributed with mean $1/\mu$. Since all customers join a single waiting queue, the server can only recognize the type of the customer at the head of the queue. The probability that the customer at the head of the queue is a Type 1 customer is p. Denote the number of customers in the system at time t by $N^2(t)$. Denote the state of server at time t by $I^2(t)$, where

$$I^2(t) = \begin{cases} 0, & \text{the server is idle at time } t, \\ 1, & \text{the server is busy with a Type 1 customer at time } t, \\ 2, & \text{the server is busy with a Type 2 customer at time } t. \end{cases} \tag{3.1}$$

The process $X^2(t) = \{(N^2(t), I^2(t)), t \geq 0\}$ is also a continuous-time Markov process with the state space $\Omega^2 = (0, 0) \cup \{(n, i), n = 0, 1, 2, \cdots, i = 1, 2\}$. The transitions between states are shown in Figure 2.

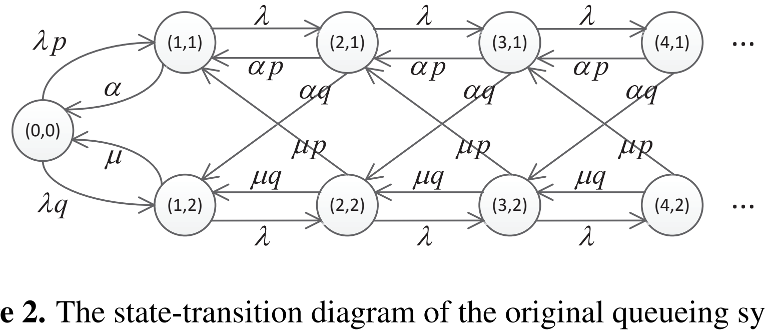


Figure 2. The state-transition diagram of the original queueing system.

The steady-state probability $p^2_{n,i}$ is defined by

$$p^2_{n,i} = \lim_{t\to\infty} Pr(N^2(t) = n, I^2(t) = i), (n, i) \in \Omega^2. \tag{3.2}$$

Arranging the Markov process in lexicography sequence, we obtain the following infinitesimal generator:

$$G = \begin{pmatrix} M_0 & N_0 & & & \\ L_0 & M & N & & \\ & L & M & N & \\ & & \ddots & \ddots & \ddots \end{pmatrix}, \tag{3.3}$$

where

$$L = \begin{pmatrix} \alpha p & \alpha q \\ \mu p & \mu q \end{pmatrix}, \quad M = \begin{pmatrix} -(\lambda+\alpha)p & 0 \\ 0 & -(\lambda+\mu) \end{pmatrix}, \quad N = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}, \tag{3.4}$$

$$M_0 = -\lambda,\ N_0 = (\lambda p \quad \lambda q),\ L_0 = (\alpha \quad \mu)'. \tag{3.5}$$

Applying the mean drift method, we can obtain the steady state condition for Model 2 as: $\lambda < \frac{\mu\alpha}{\alpha q+\mu p}$. The steady condition can be rewritten as $\frac{1}{\lambda} > \frac{q}{\mu} + \frac{p}{\alpha}$, which can be interpreted as the arrival interval $(\frac{1}{\lambda})$ is greater than the average service time $(\frac{q}{\mu} + \frac{p}{\alpha})$. The balance equations between the above states can be written as follows:

$$\lambda p^2_{0,0} = \mu p^2_{1,2} + \alpha p^2_{1,1}, \tag{3.6}$$

$$(\lambda+\alpha)p^2_{1,1} = \lambda p p^2_{0,0} + \mu p p^2_{2,2} + \alpha p p^2_{2,1}, \tag{3.7}$$

$$(\lambda+\mu)p^2_{1,2} = \lambda q p^2_{0,0} + \mu q p^2_{2,2} + \alpha q p^2_{2,1}, \tag{3.8}$$

$$(\lambda+\alpha)p^2_{n,1} = \lambda p^2_{n-1,1} + \mu p p^2_{n+1,2} + \alpha p p^2_{n+1,1}, \quad n \geq 2, \tag{3.9}$$

$$(\lambda+\mu)p^2_{n,2} = \lambda p^2_{n-1,2} + \mu q p^2_{n+1,2} + \alpha q p^2_{n+1,1}, \quad n \geq 2. \tag{3.10}$$

Define the partial generating function of the states as: $G_i(z) = \sum_{n=1}^{\infty} p^2_{n,i} z^n, z \in (0,1) \quad i = 1, 2$. By simple algebraic manipulation, the partial generating function $G_i(z)$ have the following expression:

$$G_1(z) = -p^2_{0,0} \cdot \frac{\lambda p(1-z)z((\lambda+\mu)z - \lambda z^2)}{(-\lambda z^2 + (\lambda+\mu)z - \mu q)(-\lambda z^2 + (\lambda+\alpha)z - \alpha p) - \alpha\mu pq}, \tag{3.11}$$

$$G_2(z) = -p^2_{0,0} \cdot \frac{\lambda q(1-z)z((\lambda+\alpha)z - \lambda z^2)}{(-\lambda z^2 + (\lambda+\mu)z - \mu q)(-\lambda z^2 + (\lambda+\alpha)z - \alpha p) - \alpha\mu pq}. \tag{3.12}$$

Combining with the nominal condition $G_1(1) + G_2(1) + p^2_{0,0} = 1$, the probability that the system is idle without customers can be obtained as

$$p^2_{0,0} = \frac{\mu\alpha - \lambda(\alpha q + \mu p)}{\mu\alpha}. \tag{3.13}$$

In order to compare the original system with Model 1, it is also necessary to give the expression of the expected queue length. According to the properties of the generating function, the number of customers in the system can be expressed as

$$E(N_2) = \sum_{i=1}^{2}\sum_{n=1}^{\infty} n p^2_{n,i} = G_1^{'}(1) + G_2^{'}(1) = \frac{\lambda[\lambda pq(\alpha-\mu)^2 + \alpha\mu(\alpha q + \mu p)]}{\alpha\mu(\alpha\mu - \lambda(\alpha q + \mu p))}. \tag{3.14}$$

3.1. The effect of service decomposition

In this part, we mainly focus on analyzing the effect of service decomposition among Models 1&2. We should emphasize that both the stationary condition of the two models should be satisfied $\lambda < min(\frac{\alpha\beta}{q\alpha+\beta}, \frac{\alpha\mu}{q\alpha+p\mu})$. Another condition that should be provided is $\alpha > \mu$. Otherwise, service decomposition has only negative effects. Under the above two conditions, we compare the cost function of Models 1&2 and give the condition under which the manager prefers to adopt the idea of service decomposition.

3.1.1. Cost function

Two cost components are considered: the waiting cost rate of customers in the system, c and the inventory holding cost rate, h. It is evident that the Model 2 does not have the inventory holding cost. The total expected cost of the queueing-inventory system (Model 1) and the queueing system (Model 2) are denoted by C_1 and C_2

$$C_1 = cE(N_1) + hE(I_1), \tag{3.15}$$

$$C_2 = cE(N_2). \tag{3.16}$$

Proposition 3.1. *Considering the above two cost components, management with Model 1 is more profitable if and only if the following condition is satisfied*

$$\kappa_1 \equiv \frac{h}{c} < \frac{E(N_2) - E(N_1)}{E(I_1)} = \frac{(\alpha-\mu)((\alpha-\lambda)(\alpha-\mu)\lambda p + \alpha^2\mu)(\alpha\beta - \lambda(\alpha q + \beta))}{\alpha\mu(\alpha\mu - \lambda(\alpha q + \mu p))(\lambda\beta + \alpha^2 - \lambda\alpha p)} \equiv \Delta_1. \tag{3.17}$$

From Proposition 3.1, if $\Delta_1 > 0$, we can conclude that applying service decomposition to Type 2 customers reduces the queue length but simultaneously increases the holding cost of inventory orders. Moreover, whether to adopt the service decomposition strategy proposed in Model 1 depends only on Eq (3.17). We further examine the improvement achieved by Model 1 compared with Model 2. The improvement is evaluated in terms of the percentage reduction in the total expected cost and is given by

$$\epsilon_1 = \frac{C_2 - C_1}{C_2} = \frac{cE(N_2) - cE(N_1) - hE(I_1)}{cE(N_2)}. \tag{3.18}$$

Proposition 3.2. *The percentage reduction ϵ_1 in the total cost of Model 1 can be written in the following form:*

$$\epsilon_1 = (1 - \frac{\alpha\mu[\alpha\mu - \lambda(q\alpha + p\mu)]}{(\alpha-\lambda)[\lambda pq(\alpha-\mu)^2 + \alpha\mu(\alpha q + \mu p)]})(1 - \frac{\kappa_1}{\Delta_1}).$$

Proof.

$$\begin{aligned}
\epsilon_1 &= \frac{E(N_2) - E(N_1)}{E(N_2)} - \frac{h}{c}\frac{E(I_1)}{E(N_2)} = \frac{E(N_2) - E(N_1)}{E(I_1)}\frac{E(I_1)}{E(N_2)} - \kappa_1\frac{E(I_1)}{E(N_2)} \\
&= \frac{E(I_1)}{E(N_2)}(\Delta_1 - \kappa_1) = \Delta_1\frac{E(I_1)}{E(N_2)}(1 - \frac{\kappa_1}{\Delta_1}) = \frac{E(N_2) - E(N_1)}{E(N_2)}(1 - \frac{\kappa_1}{\Delta_1}) \\
&= (1 - \frac{\alpha\mu[\alpha\mu - \lambda(q\alpha + p\mu)]}{(\alpha-\lambda)[\lambda pq(\alpha-\mu)^2 + \alpha\mu(\alpha q + \mu p)]})(1 - \frac{\kappa_1}{\Delta_1})
\end{aligned}$$

□

To investigate the influence of the system's parameters on model selection, we give the trend of Δ_1 in some extreme cases.

Corollary 3.1. *(1) as* $\lambda \to 0$, $\Delta_1 \to \frac{\beta(\alpha-\mu)}{\alpha\mu}$*; (2) when* $\frac{1}{\mu} < \frac{1}{\alpha} + \frac{1}{\beta}$*, as* $\lambda \to \frac{\alpha\beta}{\alpha q+\beta}^{-}$, $\Delta_1 \to 0$*; (3) when* $\frac{1}{\mu} > \frac{1}{\alpha} + \frac{1}{\beta}$*, as* $\lambda \to \frac{\alpha\mu}{\alpha q+\mu p}^{-}$, $\Delta_1 \to \infty$.

Proof. Under the assumption $\alpha > \max(\mu, \lambda)$, $(\alpha - \mu)$, $[(\alpha - \lambda)(\alpha - \mu)\lambda + \alpha^2\mu]$ and $(\lambda\beta + \alpha^2 - \lambda\alpha p)$ in the expression of Δ_1 are always positive.

(1) This conclusion is obvious when we substitute $\lambda \to 0$ into the expression for Δ_1 in Eq (3.17). It can be verified in Figure 3(a) and 3(b).

(2) When $\frac{1}{\mu} < \frac{1}{\alpha} + \frac{1}{\beta}$, we can conclude that $\frac{\alpha\beta}{\alpha q+\beta} < \frac{\alpha\mu}{\alpha q+\mu p}$. Then if $\lambda \to \frac{\alpha\beta}{\alpha q+\beta}$, $(\alpha\mu - \lambda(\alpha q + \mu p))$ is positive and $\Delta_1 \to 0$.

(3) When $\frac{1}{\mu} > \frac{1}{\alpha} + \frac{1}{\beta}$, we can conclude that $\frac{\alpha\beta}{\alpha q+\beta} > \frac{\alpha\mu}{\alpha q+\mu p}$. Then if $\lambda \to \frac{\alpha\mu}{\alpha q+\mu p}$, $(\alpha\beta - \lambda(\alpha q + \beta))$ is positive and $\Delta_1 \to \infty$. □

Remark 3. Corollary 3.1 gives the values of Δ_1 for the three extreme cases. (1)When the customer arrival rate λ is low($\lambda \to 0^+$), Δ_1 tends to a constant with fixed α, β and μ. This suggests that the manager's decision, in this case, depends mainly on the relationship between $\frac{h}{c}$ and Δ_1. If Model 1 is more profitable, the percentage reduction in the total cost is given by $\epsilon_1 \overset{\lambda\to 0^+}{\longrightarrow} \frac{q(\beta(\alpha-\mu)-\kappa_1\alpha\mu)}{\beta(\alpha q+\mu p)}$. (2) When $\frac{1}{\mu} < \frac{1}{\alpha} + \frac{1}{\beta}$, the average full service time for Type 2 customers is smaller than the total average decomposed service time. Model 2 is beneficial as $\lambda \to \frac{\alpha\beta}{\alpha q+\beta}^{-}$ regardless of the ratio of c,h; see Figure 3(a). This is in line with our intuitive guess. In Model 1, when the customer arrival rate is high ($\lambda \to \frac{\alpha\beta}{\alpha q+\beta}^{-}$), there are two queues in the system. One is the customers' waiting queue, and the other is for stored orders. Service decomposition reduces customers' waiting time while increasing the cost of maintaining the stored orders. Conversely, the increased sojourn time of Type 2 customers is less than the processing time of stored orders. Then Model 2 is preferable even for a low but positive h. See Figure 3(a). (3) When $\frac{1}{\mu} > \frac{1}{\alpha} + \frac{1}{\beta}$, the total average decomposed service time is shorter than the average continuous service time. If the customer arrival rate is high ($\lambda \to \frac{\alpha\mu}{\alpha q+\mu p}^{-}$), the customer average sojourn time is high. Service decomposition can largely reduce the queue length and lower costs even for a large but finite h. In this case, the economic improvement is $\epsilon_1 \overset{\lambda\to \frac{\alpha\mu}{\alpha q+\mu p}^{-}}{\longrightarrow} 1$. See Figure 3(b).

Corollary 3.2. *(1) As* $\beta \to \frac{\lambda\alpha q}{\alpha-\lambda}^{+}$, $\Delta_1 \to 0$*; (2) As* $\beta \to \infty$, $\Delta_1 \to \frac{(\alpha-\mu)\big((\alpha-\lambda)(\alpha-\mu)\lambda p+\alpha^2\mu\big)(\alpha-\lambda)}{\alpha\mu\lambda\big(\alpha\mu-\lambda(\alpha q+\mu p)\big)}$.

Proof. Straightforward by substituting in Eq (3.17) and rearranging items. □

Remark 4. Corollary 3.2 indicates that Δ_1 tends to zeros for a low additional service rate ($\beta \to \frac{\lambda\alpha q}{\alpha-\lambda}^{+}$). This can help managers make quick decisions on model selection: if the additional service β rate is less than $\frac{\lambda\alpha q}{\alpha-\lambda}$, retain the full service for Type 2 customer instead of service decomposition. See Figure 3(c). When the system processes pending orders in inventory at a high rate ($\beta \to \infty$), a new arrival can preempt the current service. In this case, Δ_1 tends to be a positive constant. This implies that the manager may retain the full service of Type 2 customers (Model 2) even for a high additional service rate. See Figure 3(c).

Corollary 3.3. *(1) As* $\alpha \to \max(\mu, \frac{\lambda\beta}{\beta-\lambda q})^{+}$, $\Delta_1 \to 0$*; (2) As* $\alpha \to \frac{\lambda\mu p}{\mu-\lambda q}$, $\Delta_1 \to \infty$*;(3) As* $\alpha \to \infty$, $\Delta_1 \to \frac{(\lambda p+\mu)(\beta-\lambda q)}{\mu(\mu-\lambda q)}$.

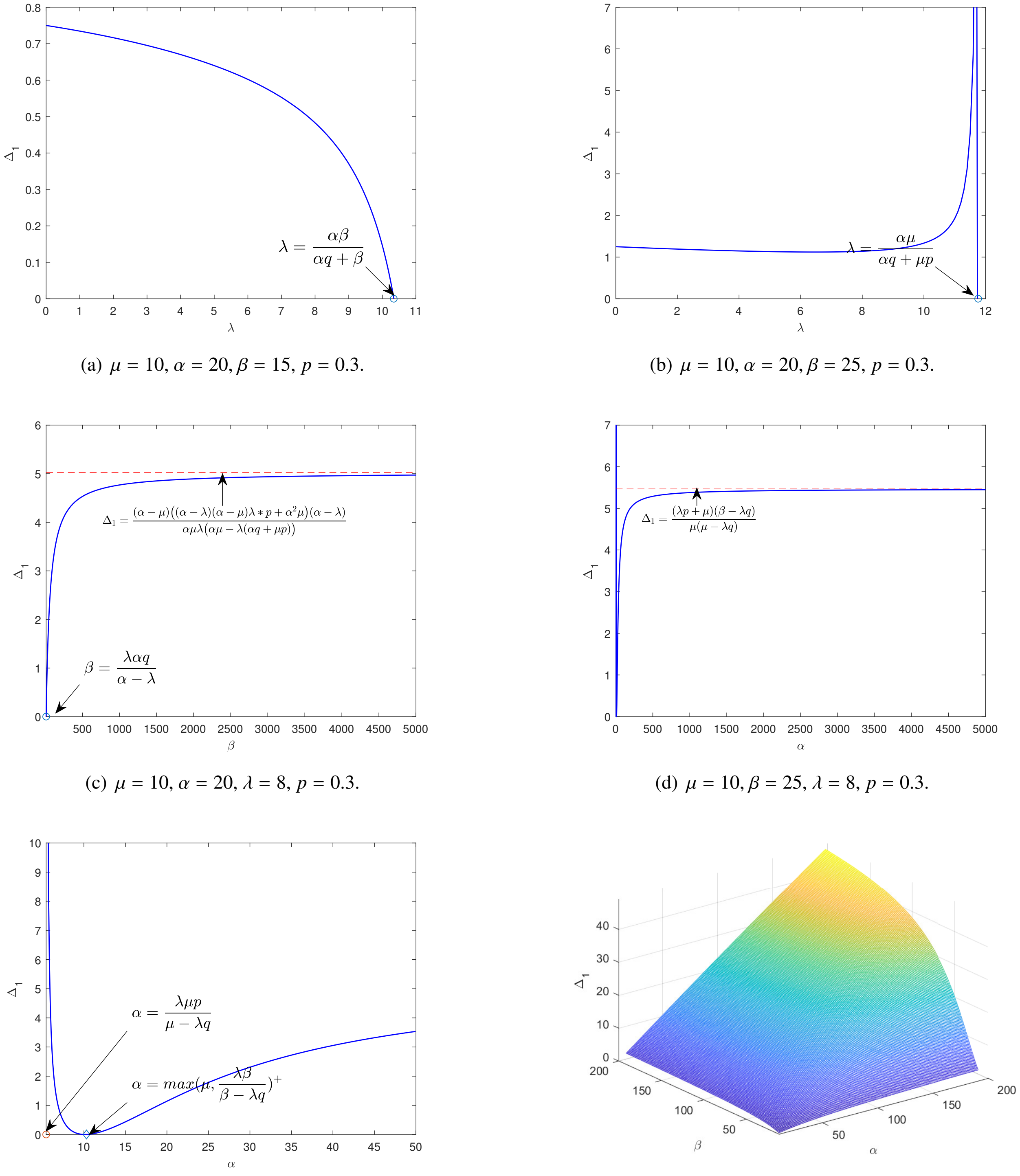
Proof. Straightforward by substituting in Equation (3.17) and rearranging items. □

Remark 5. Corollary 3.3 summarizes the values of Δ_1 for extreme values of the basic service rate α. In the first case, Δ_1 tends to zeros when the basic service rate α is small ($\alpha \to \max(\mu, \frac{\lambda\beta}{\beta-\lambda q})^+$). This situation is consistent with our intuitive inference. If the basic service rate α gradually approaches the full service rate μ for Type 2 customers, the system will process pending orders in inventory during idle time. This situation does not reduce the customer's waiting cost while increasing the system's inventory holding cost. Therefore Model 2 is more profitable than Model 1; see Figure 3(e). In the second case, Δ_1 tends to be infinite when the basic service rate is small $\alpha \to \frac{\lambda\mu p}{\mu-\lambda q}$. It seems that Model 1 is more appropriate in this case. See Figure 3(e). However, this occurs only when the basic service rate is smaller than the full service rate of Type 2 customers ($\alpha < \mu$). This contradicts our previous assumptions ($\alpha > \mu$), so we exclude this extreme case in the following discussion. In the last case, for a large basic service rate ($\alpha \to \infty$), Δ_1 tends to a finite constant. This means that Model 2 is preferable regardless of the ratio of h and c. If the basic service rate α is large, only the pending orders queue in the inventory. Compared to Model 2, Model 1 is more profitable when the cost of Model 1 to hold inventory is less than the customers' waiting cost in Model 2. Also known as $\kappa_1 < \Delta_1 = \frac{(\lambda p+\mu)(\beta-\lambda q)}{\mu(\mu-\lambda q)}$. Regarding economic improvement, the percentage reduction in the total cost of this case can be obtained by $\epsilon_1 = 1 - \kappa_1 \frac{\mu(\mu-\lambda q)}{(\lambda p+\mu)(\beta-\lambda q)}$. Figure 3(d) illustrates the trend of Δ_1.

Corollary 3.4. *(1) As* $\{\alpha,\beta\} \to \{\mu^+, \infty\}$, $\Delta_1 \to 0$*; (2) As* $\{\alpha,\beta\} \to \{\infty, \lambda q^+\}$, $\Delta_1 \to 0$*; (3) As* $\{\alpha,\beta\} \to \{\infty, \infty\}$, $\Delta_1 \to \infty$.

Proof. Straightforward by substituting in Eq (3.17) and rearranging items. □

Remark 6. Corollary 3.4 shows that Model 2 is preferable if the basic service rate α is small ($\alpha \to \mu^+$), even for a high additional service rate ($\beta \to \infty$). It is easy to explain that as $\alpha \to \mu$, the service rates of the two types of customers gradually equalize. It would only increase the cost of the system to split the Type 2 customer's service into two phases, even if the additional service rate is infinite. The same outcome holds when the additional service rate β is low ($\beta \to \lambda q^+$), even for a large basic service rate ($\alpha \to \infty$). This result is intuitive. When the additional service rate is low, many pending orders accumulate in the inventory, significantly increasing the system holding cost. Conversely, when the basic service rate α and the additional service rate β are significant, Δ_1 tends to infinity. At this point, Model 1 is beneficial for almost all values of c and h. The economic improvement for the case with high values of the basic service rate and the additional service rate ($\{\alpha,\beta\} \to \{\infty, \infty\}$) is given by $\epsilon_1 \to 1$. Figure 3(f) shows the trends of the values of Δ_1.



(a) $\mu = 10, \alpha = 20, \beta = 15, p = 0.3$.

(b) $\mu = 10, \alpha = 20, \beta = 25, p = 0.3$.

(c) $\mu = 10, \alpha = 20, \lambda = 8, p = 0.3$.

(d) $\mu = 10, \beta = 25, \lambda = 8, p = 0.3$.

(e) Partial enlarged view of Figure 3(d).

(f) $\lambda = 8, \mu = 10$.

Figure 3. The effects of different parameters on Δ_1.

3.1.2. System idle time

We now examine the impact of service decomposition on system idle time. The idle probability can be used to represent the average idle fraction. In Model 1, the server is dormant only when there are no customers in the system and no stored orders in the inventory. Thus, the idle probability of Model 1 is $p^1_{0,0}$. The idle probability of Model 2 is $p^2_{0,0}$. Obviously, we can obtain the following proposition.

Proposition 3.3. *The idle fraction of Model 1 to be larger than Model 2, if* $\frac{1}{\mu} > \frac{1}{\alpha} + \frac{1}{\beta}$ *is satisfied.*

Proof. Directly substitute the two probability expressions to complete the proof.

$$p^1_{0,0} = \frac{(\alpha - \lambda)\beta - \lambda\alpha q}{\alpha\beta} > \frac{\alpha\mu - \lambda(\alpha q + \mu p)}{\alpha\mu} = p^2_{0,0}$$

$$\mu((\alpha - \lambda)\beta - \lambda\alpha q) > \beta(\alpha\mu - \lambda(\alpha q + \mu p))\frac{1}{\mu} > \frac{1}{\alpha} + \frac{1}{\beta}$$

□

Remark 7. The total time of decomposed services is generally longer than the average time of continuous complete services $\frac{1}{\mu} < \frac{1}{\alpha} + \frac{1}{\beta}$, due to the additional work resulting from service decomposition: e.g., allocation management of pending orders, confirmation of order information, etc. However, in some cases the opposite ($\frac{1}{\mu} > \frac{1}{\alpha} + \frac{1}{\beta}$) also exists. It is also inevitable that customers interfere with the service process during service. It can lead to a lower actual service rate. In addition, when performing additional service, the server can process similar orders in bulk, increasing the additional service rate. Proposition 3.3 implies that under condition $\frac{1}{\mu} > \frac{1}{\alpha} + \frac{1}{\beta}$, service decomposition can increase the idle time of the system. Using system idle time to process orders in inventory does not reduce the idle fraction of the system, which seems contrary to our common sense.

4. Numerical results

4.1. Model 3: the queueing-inventory system with finite capacity

In Remark 2, we mentioned that a Type 2 customer might request a full service due to the excessive waiting time for a stored order. The system inventory level would then be a finite value, which is also more in line with the actual situation. In this subsection, we focus on another queueing-inventory model in which the service of a Type 2 customer is split into two parts only when the inventory level is less than N. That means the inventory capacity is limited and only allows a maximum of N orders to be stored. If the inventory level is less than N, the system operates as Model 1. Conversely, if the inventory capacity is N, a Type 2 customer will receive a full service. At this point, the system operates in line with Model 2. Therefore, Model 3 can be seen as a mixture of Models 1&2.

Denote the number of customers in the system, the server's state and the inventory level at time t by $N^3(t)$, $S(t)$ and $I^3(t)$, respectively, where

$$S(t) = \begin{cases} 0, & \text{the server is idle,} \\ 1, & \text{the server is busy with a Type 1 customer,} \\ 2, & \text{the server is busy with a Type 2 customer,} \\ 3, & \text{the server is in the basic service,} \\ 4, & \text{the server is in the additional service.} \end{cases} \tag{4.1}$$

The state space $\Omega^3 = \cup_{n=0}^{\infty}\{\boldsymbol{H_n}\}$, where

$$\begin{aligned} \boldsymbol{H_0} &= \{(0,0,0),(0,4,1),(0,4,2),\cdots,(0,4,N)\},\\ \{\boldsymbol{H_n}\}_{n\geq 1} &= \{(n,3,0),(n,3,1),(n,3,2),\cdots,(n,3,N-1),(n,1,N),(n,2,N)\}. \end{aligned}$$

The transitions between all the states can be illustrated in Figure 4.

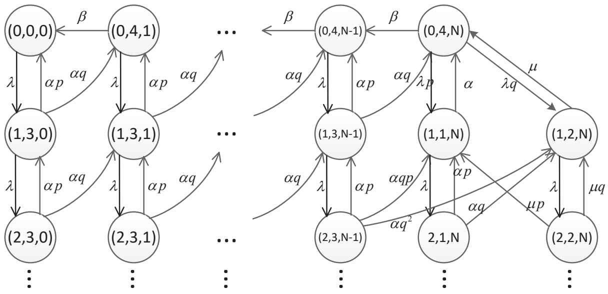


Figure 4. The state-transition diagram of Model 3.

Arranging the Markov process $X^3(t) = \{(N^3(t), S(t), I^3(t)), t \geq 0\}$ in lexicography sequence, we can obtain the following infinitesimal generator:

$$Q = \begin{pmatrix} B_{00} & C_{01} & & & \\ A_{10} & B & C & & \\ & A & B & C & \\ & & A & B & C \\ & & & \ddots & \ddots & \ddots \end{pmatrix}, \tag{4.2}$$

where

$$B_{00} = \begin{pmatrix} -\lambda & & & & \\ \beta & -(\lambda+\beta) & & & \\ & \beta & -(\lambda+\beta) & & \\ & & \ddots & \ddots & \\ & & & \beta & -(\lambda+\beta) \end{pmatrix}_{(N+1)\times(N+1)}, C_{01} = \begin{pmatrix} \lambda & & & & & \\ & \lambda & & & & \\ & & \ddots & & & \\ & & & \lambda & & \\ & & & & \lambda p & \lambda q \end{pmatrix}_{(N+1)\times(N+2)}, \tag{4.3}$$

$$A_{10} = \begin{pmatrix} \alpha p & \alpha q & & & \\ & \alpha p & \alpha q & & \\ & & \ddots & \ddots & \\ & & & \alpha p & \alpha q \\ & & & & \alpha \\ & & & & \mu \end{pmatrix}_{(N+2)\times(N+1)}, A = \begin{pmatrix} \alpha p & \alpha q & & & & & \\ & \alpha p & \alpha q & & & & \\ & & \ddots & \ddots & & & \\ & & & \alpha p & \alpha q & & \\ & & & & \alpha p & \alpha pq & \alpha q^2 \\ & & & & & \alpha p & \alpha q \\ & & & & & \mu p & \mu q \end{pmatrix}_{(N+2)\times(N+2)}, \tag{4.4}$$

and

$$B = diag(-(\lambda+\alpha), -(\lambda+\alpha), \cdots, -(\lambda+\alpha), -(\lambda+\mu)), \quad C = diag(\lambda, \lambda, \cdots, \lambda). \tag{4.5}$$

B and C are both square arrays of order $N+2$. The steady condition for Model 3 can be obtained by the mean drift method. Let $\boldsymbol{\xi} = (\xi_0, \xi_1, \cdots, \xi_{N-1}, \xi_{N,1}, \xi_{N,2})$ be te invariant probability vector of the matrix:

$$D = A + B + C = \begin{pmatrix} -\alpha q & \alpha q & & & & & \\ & -\alpha q & \alpha q & & & & \\ & & \ddots & \ddots & & & \\ & & & -\alpha q & \alpha q & & \\ & & & & -\alpha q & \alpha pq & \alpha q^2 \\ & & & & & -\alpha q & \alpha q \\ & & & & & \mu p & -\mu p \end{pmatrix}. \tag{4.6}$$

We can obtain vector $\boldsymbol{\xi}$ from $\boldsymbol{\xi} D = 0$ and $\boldsymbol{\xi e} = 1$ ($\boldsymbol{e}$ represents the all-1 vector of the corresponding order) as:

$$\begin{cases} \xi_k = 0, \quad k = 0, 1, \cdots, N-1, \\ \xi_{N,1} = \dfrac{\mu p}{\alpha q + \mu p}, \\ \xi_{N,2} = \dfrac{\alpha q}{\alpha q + \mu p}. \end{cases} \tag{4.7}$$

According to the mean-drift method, the stationary condition of Model 3 is $\boldsymbol{\xi A e} > \boldsymbol{\xi C e}$, which can be rewritten as $\lambda < \frac{\alpha\mu}{\alpha q + \mu p}$. When the number of customers in the system accumulates to a certain number, the inventory level will reach the maximum capacity N. After that, the service rules of system customers are the same as Model 2, so the steady-state conditions are also the same.

Next, we analyze this queueing-inventory system with limited capacity under the stability condition. The Matrix-Geometric method developed by Neuts [19] is applied to derive the stationary distribution of Model 3. First, we define the stationary joint probability $p^3_{n,s,i}$ as

$$\begin{aligned} p^3_{n,s,i} &= \lim_{t\to\infty} Pr[N^3(t) = n, S(t) = s, I^3(t) = i], \quad (n, s, i) \in \Omega^3, \\ p^3_0 &= (p^3_{0,0,0},\ p^3_{0,4,1}, \cdots,\ p^3_{0,4,N}) \\ p^3_n &= (p^3_{n,3,0},\ p^3_{n,3,1}, \cdots,\ p^3_{n,3,N-1},\ p^3_{n,1,N},\ p^3_{n,2,N}),\ n \geq 1 \\ p^3 &= (p^3_0,\ p^3_1, ; \cdots,\ p^3_n, \cdots) \end{aligned} \tag{4.8}$$

We should emphasize that the stationary joint probability p^3 must satisfy the following equations: $p^3 Q = 0$ and $p^3 \boldsymbol{e} = 1$, which equal to

$$p^3_0 B_{00} + p^3_1 A_{01} = 0, \tag{4.9}$$

$$p^3_0 C_{01} + p^3_1 B + p^3_2 A = 0, \tag{4.10}$$

$$p^3_{n-1} C + p^3_n B + p^3_{n+1} A = 0, \quad n \geq 2. \tag{4.11}$$

If we assume that the stationary p^3_n have the following expression:

$$p^3_n = p^3_1 R^{n-1},\ n \geq 2, \tag{4.12}$$

where R is a matrix of order $(N+2)\times(N+2)$ satisfying $C+RB+R^2A=0$. The boundary probability vectors p_0^3, p_1^3 can be determined by Eqs (4.9) and (4.10) and the normalizing condition as follows:

$$\begin{cases} p_0^3B_{00}+p_1^3A_{01}=0,\\ p_0^3C_{01}+p_1^3(B+RA)=0,\\ p_0^3\boldsymbol{e}+p_1^3(\boldsymbol{I}-R)^{-1}\boldsymbol{e}=1. \end{cases} \tag{4.13}$$

Since R cannot be computed explicitly, we can obtain a numerical result for R with the algorithms devised by Latouche and Ramaswami [20]. Then, we list two performance measures of Model 3.

(1) The average number of customers in the system is denoted by L:

$$L=\sum_{n=0}^{\infty}np_n^3\boldsymbol{e}=p_1^3[I-R]^{-2}\boldsymbol{e}; \tag{4.14}$$

(2) Let I_q denote the mean number of stored orders in inventory. We have

$$I_q=p_0^3\boldsymbol{v}+\sum_{i=1}^{\infty}p_i^3\boldsymbol{w}=p_0^3\boldsymbol{v}+p_1^3(\sum_{i=1}^{\infty}R^{i-1})\boldsymbol{w}=p_0^3\boldsymbol{v}+p_1^3[I-R]^{-1}\boldsymbol{w}, \tag{4.15}$$

where $\boldsymbol{v}=(0,0,1,2,\cdots,N-1)^T$ and $\boldsymbol{w}=(0,1,2,3,\cdots,N-1,N,N)^T$.

4.2. Managerial implications

In this subsection, we consider determining the optimal inventory capacity N to reduce system costs under different parameter settings. The performance measures can be expressed as a function of the maximum inventory capacity N, i.e., $L(N)$ and $I_q(N)$. First, we should establish a total cost function for the queueing-inventory system with limited capacity as

$$Tc(N)=cL(N)+hI_q(N), \tag{4.16}$$

where c and h are defined as Section 2. Without loss of generality, we set the customer's sojourn time cost rate to 1, i.e., $c=1$. Also, set h to 0.6. This means that the unit cost of maintaining an inventory order costs is less than that of making a customer wait for service.

Example 1. According to the previous discussion in Section 2, the relationship between $\frac{1}{\mu}$ and $\frac{1}{\alpha}+\frac{1}{\beta}$ will affect the numerical results, so we discuss the effect of N on the total cost function $Tc(N)$ in two cases. The curves of case 1 :$\frac{1}{\mu}<\frac{1}{\alpha}+\frac{1}{\beta}$ and case 2: $\frac{1}{\mu}>\frac{1}{\alpha}+\frac{1}{\beta}$ are displayed in Figure 5. The figure indicates that when case 2 holds, the total cost function decreases as N increases. This implies that performing service decomposition for all Type 2 customers is profitable. According to our intuitive conjecture, the queue length $L(N)$ decreases as N increases, which is consistent with the numerical results in Figure 6. However, the trend of the average inventory level $I_q(N)$ against N in Figure 6 is opposite to our conjecture, which can be explained as follows. When N tends to infinity, the system can store sufficient orders. Therefore, the average queue length $L(N)$ gradually tends to a constant value (queue length of the classical M/M/1 queueing system $\frac{\lambda}{\alpha-\lambda}$). However, the system inventory level does not keep growing but tends to stabilize. According to Proposition 3.3, the idle fraction is increased due to service decomposition in case 2. The customer leaves the system immediately after receiving the basic service. Thus, the pending orders in the inventory can be processed during the idle time, so the system inventory level does not keep growing. Therefore, in case 2, performing service decomposition for all Type 2 customers is beneficial. As for case 1, the blue curve in Fig. 5 demonstrates the variation of the cost function as N increases. The point (4, 2.1146) is the optimal value point, indicating that the minimum total cost can be obtained when the maximum inventory capacity of the system is set to 4 at this parameter setting. According to the previous analysis, in case 1, an optimal solution of the cost function $Tc(N)$ with respect to N exists. A simple algorithm can be used to search the optimal values over a small set of integer values of the maximum inventory level N. This suggests that, in some cases, providing service decomposition for all Type 2 customers is not optimal. Managers can set an optimal inventory capacity to obtain the optimal cost.

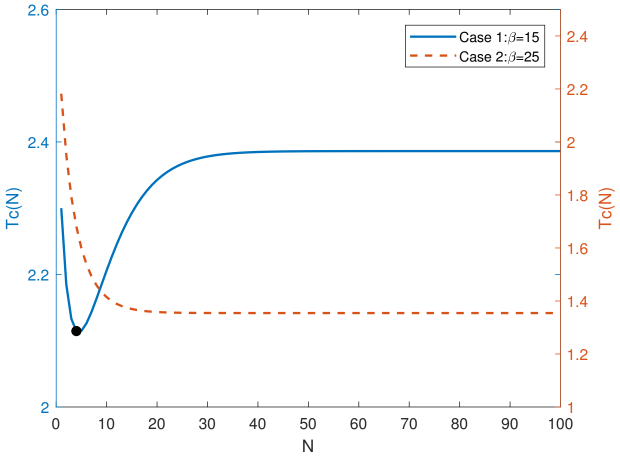


Figure 5. Cost function $Tc(N)$ as a function of N, for $\lambda = 8$, $\mu = 10$, $\alpha = 20$, p=0.3.

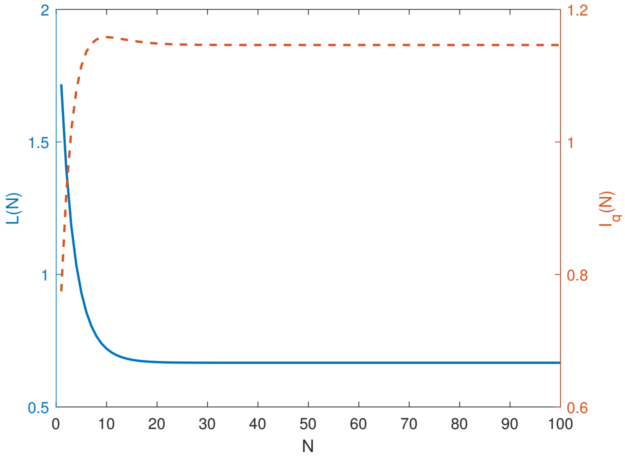


Figure 6. $L(N)\&I_q(N)$ as functions for case 2: $\lambda = 8$, $\mu = 10$, $\alpha = 20, \beta = 25$, and p=0.3.

Example 2. In this numerical example, we focus on the effect of some parameters on the optimal values N^* and $Tc^*(N)$.

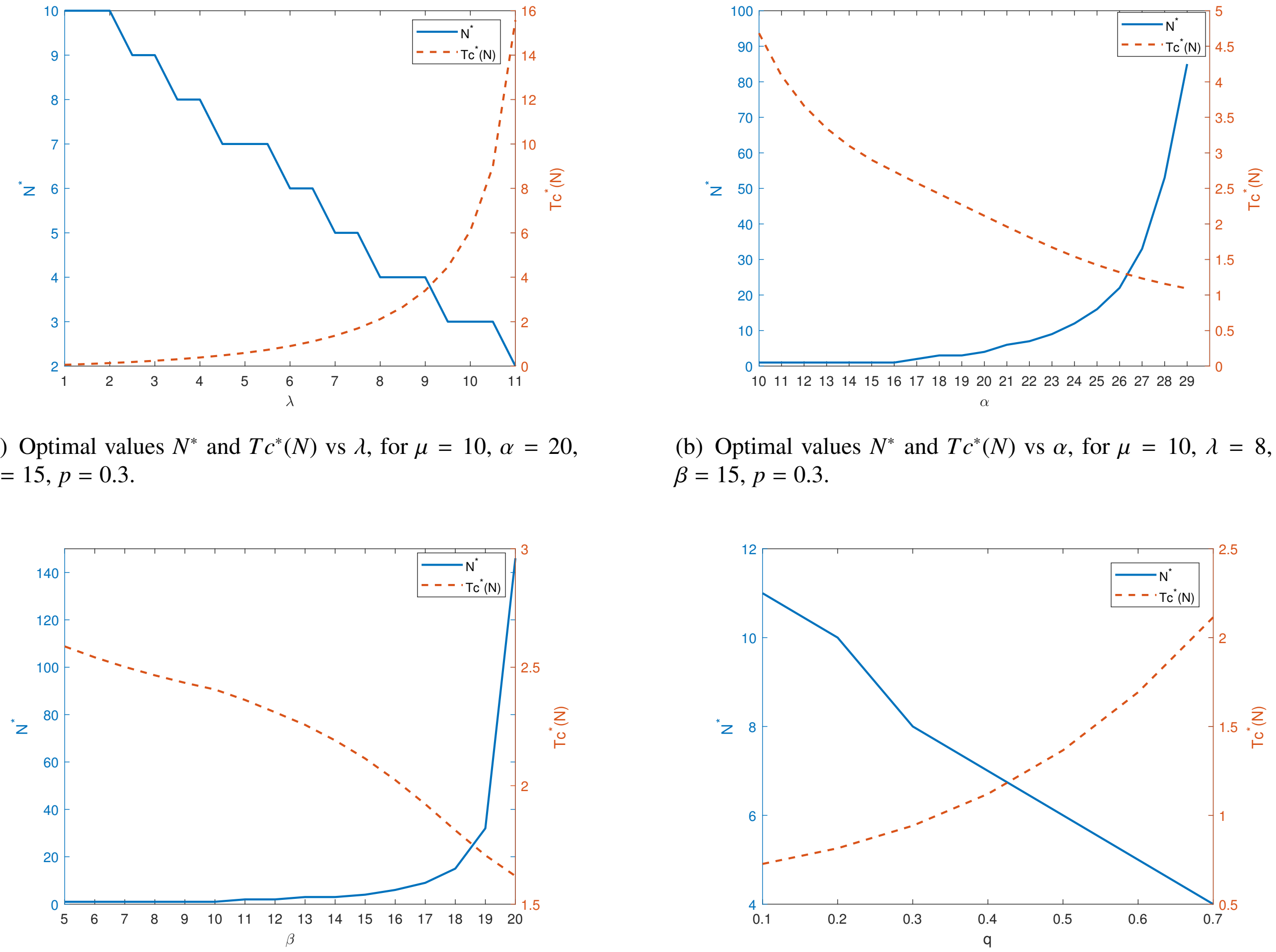
(1) Figure 7(a) shows that the system optimal inventory capacity N^* decreases as λ increases. This result contradicts our intuitive guesses, which can be explained as follows. An increase in λ leads to an increase in the number of customers arriving in the system. After the system inventory reaches a set level (maximum inventory capacity), it operates in line with Model 2. This results in the inventory level not decreasing while the queue length increases. As a result, the system cost will increase rapidly, so that N^* is reduced to reduce the inventory maintenance cost. As for the system, total cost $Tc^*(N)$ increases with the increase of λ. When $\lambda > 8$, the growth rate of $Tc^*(N)$ also becomes larger. This is because the system is gradually saturated if $\lambda > 8$. A slight increase in λ leads to a significant increase in the queue length and the total cost $Tc^*(N)$.

(2) Figure 7(b) displays the curves of N^* and $Tc^*(N)$ with respect to α. As we can see from the blue curve, the optimal inventory capacity N^* increases as α increases. When α is small, the customer queue length increases, resulting in an increase in customer waiting costs. As the service rate α increases, the number of customers waiting in the system decreases. What is more, the customer waiting cost decreases, and the system idle time increases. The orders stored in the inventory can be processed during the idle time. An increase in N^* allows the system to provide service decomposition for more Type 2 customers, thus reducing system costs which can be confirmed by the orange curve in Figure 7(b). In the previous discussion, we have assumed that the ratio of h to c is 0.6. When α increases, $Tc^*(N)$ decreases even though N^* is increasing. This implies that $cL(N)$ outperforms $hI_q(N)$.

(3) Figure 7(c) illustrates the effect of β on the optimal values $(N^*, Tc^*(N))$. The customer waiting cost is constant since β does not affect the system queue length. Managers need only consider the cost of maintaining inventory in this case. When β takes a smaller value, the average time to execute the additional service is longer. During the idle period, the additional service is more likely to be interrupted by the arrival of a new customer. Therefore, providing service decomposition for Type 2 customers is not recommended. Conversely, when $\beta > 18$, providing service decomposition to more Type 2 customers is recommended. This is because the additional service rate β is much greater than the customer arrival rate λ. During the customer arrival interval, the number of stored orders in inventory decreases even though N^* is large. The change in β only directly affects the system's inventory level. An increase in N^* means that more orders can be stored but does not represent an increase in the system's average inventory level since the inventory is not always full. Instead, the optimal inventory capacity N^* increases due to the additional service rate β increases. More Type 2 customers can leave the system immediately after receiving basic services, which reduces the waiting time for all customers. Thus, the optimal total cost $Tc^*(N)$ decreases as β increases even though N^* is increasing, implying that $cL(N)$ is outperforming $hI_q(N)$.

(4) Figure 7(d) describes the trends of N^* and $Tc^*(N)$ with respect to q. The increase in q represents an increase in the number of Type 2 customers arriving in the system. The blue curve indicates that the optimal inventory capacity N^* decreases as q increases. This result may be considered counter intuitive. Due to the increase in q, the system inventory level will reach the maximum inventory capacity more quickly. After this point, the system no longer provides service decomposition for Type 2 customers. Since the full service rate for Type 2 customers is less than that of Type 1 customers ($\mu < \alpha$), the waiting cost of Type 1 customers increases. In this way, the system

inventory is kept at a high level while the queue length increases, which leads to a higher total cost. Thus, providing service decomposition for more Type 2 customers is not beneficial, and vice versa; when q is small, if the inventory capacity is set smaller, more Type 2 customers will need to receive full service. This will increase the waiting cost for Type 1 customers. Therefore, in this case, providing service decomposition for more Type 2 customers is recommended, which means setting the inventory capacity N higher.



(a) Optimal values N^* and $Tc^*(N)$ vs λ, for $\mu = 10$, $\alpha = 20$, $\beta = 15$, $p = 0.3$.

(b) Optimal values N^* and $Tc^*(N)$ vs α, for $\mu = 10$, $\lambda = 8$, $\beta = 15$, $p = 0.3$.

(c) Optimal values N^* and $Tc^*(N)$ vs β, for $\mu = 10$, $\lambda = 8$, $\alpha = 20$, $p = 0.3$.

(d) Optimal values N^* and $Tc^*(N)$ vs q, for $\mu = 10$, $\lambda = 8$, $\alpha = 20, \beta = 15$.

Figure 7. The effect of different parameters on optimal values N^* and $Tc^*(N)$.

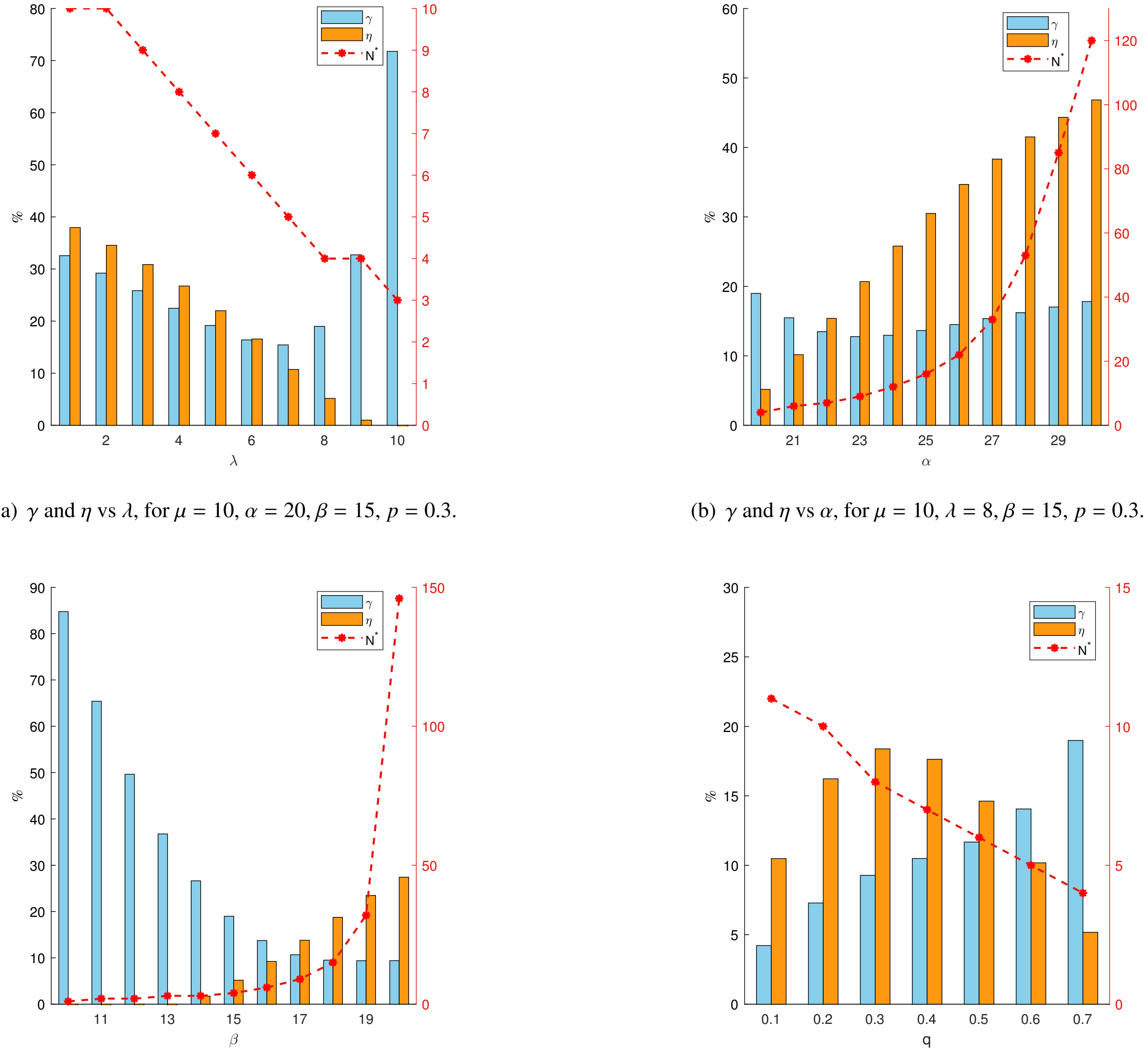
Example 3. In this example, we investigate the economic improvement of the optimal capacity setting with respect to the two extreme capacities ($n = \infty$ for Model 1 and $n = 0$ for Model 2). First, we define the improvement achieved by Model 3 as evaluated in terms of the percentage reduction in the total expected cost of Model 1 as

$$\gamma = \frac{C_1 - Tc^*(N)}{C_1}. \tag{4.17}$$

The economic improvement achieved by Model 3 is evaluated as the percentage reduction in the total average cost of Model 2, which is defined as

$$\eta = \frac{C_2 - Tc^*(N)}{C_2}. \tag{4.18}$$

Figure 8 presents the effect of system parameters on the economic percentage improvements γ and η.



(a) γ and η vs λ, for $\mu = 10$, $\alpha = 20$, $\beta = 15$, $p = 0.3$.

(b) γ and η vs α, for $\mu = 10$, $\lambda = 8$, $\beta = 15$, $p = 0.3$.

(c) γ and η vs β, for $\mu = 10$, $\lambda = 8$, $\alpha = 20$, $p = 0.3$.

(d) γ and η vs q, for $\mu = 10$, $\lambda = 8$, $\alpha = 20$, $\beta = 15$.

Figure 8. The effect of different parameters on γ and η.

(1) From Figure 8(a), we can observe that both the optimal decision N^* and the monetary improvement η (compared with Model 2) decrease with λ. As N^* decreases, Model 3 gradually converges to Model 2. So η has the same trend as N^*. γ decreases as λ increases when $\lambda < 7$.

However, once $\lambda > 7$, there is a rapid increase in γ. The former result may be the opposite of our conjecture. Nevertheless, it can be explained as follows. γ can be rewritten as $1 - \frac{Tc^*(N)}{C_1}$. When $\lambda < 7$, $Tc^*(N)$ increases at a greater rate than C_1, so γ shows a decreasing trend.

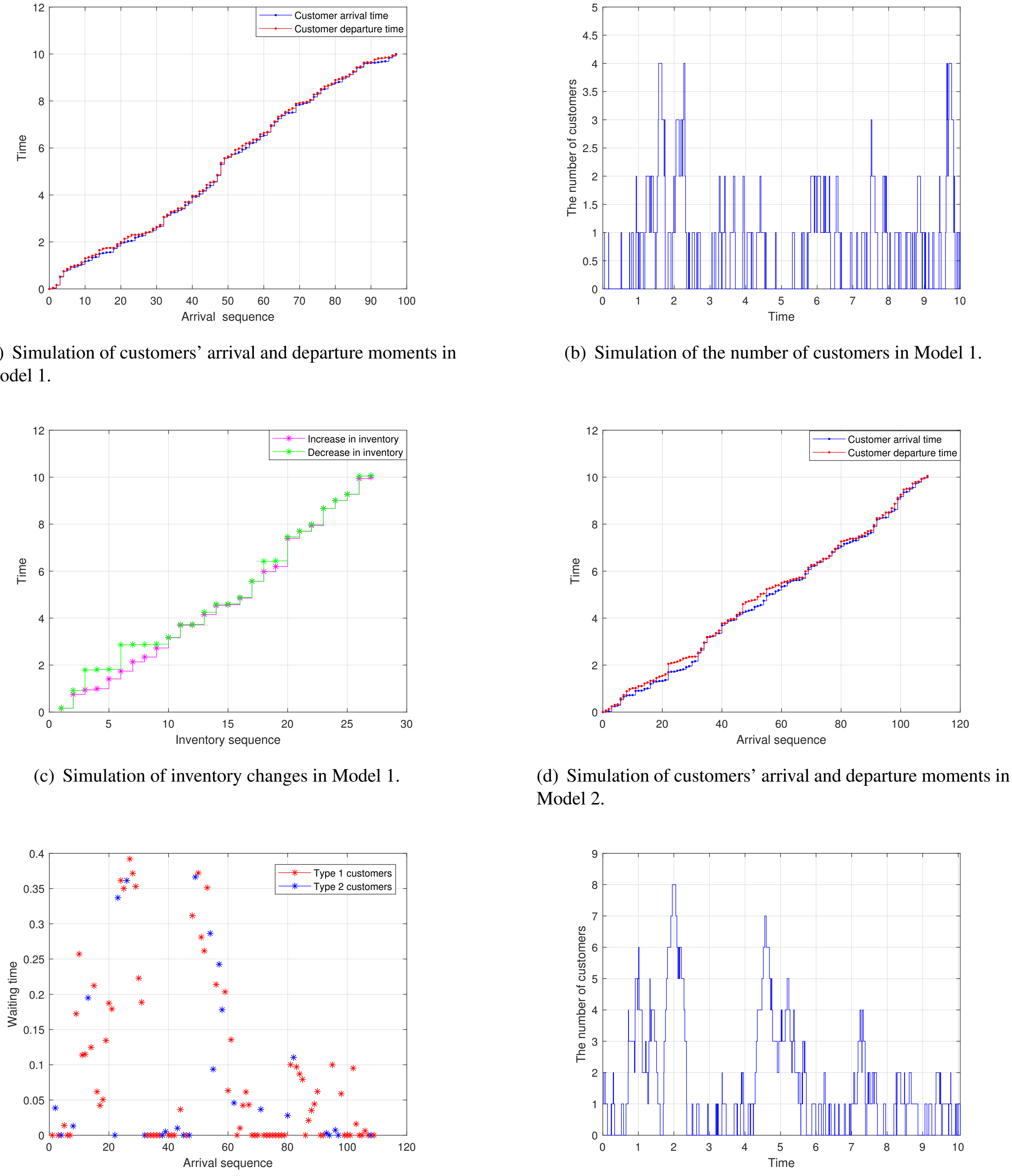
(2) Figure 8(b) displays the curves of (N^*, γ, η) against α. It can be observed that N^* increases rapidly, and η increases with α. As for γ, it shows a trend of decreasing first and then increasing with the increase of α. The increasing trend of γ can be explained as the rate at which $Tc^*(N)$ approaches C_1 is smaller than the rate at which C_1 grows, for a large α, although the N^* is large ($N^* > 80$).

(3) Figure 8(c) indicates that both the optimal decision N^* and the monetary improvement η increase with β, which are consistent with our conjectures. We notice that the economic improvement γ decreases with β. What is more, when $\beta < 13$, the percentage reduction γ is large, especially for $\beta = 10$. This result is consistent with the first case in Corollary 3.2. Therefore, managers are advised to provide only service decomposition for some Type 2 customers in this case.

(4) Figure 8(d) presents the effect of q on (γ, η, N^*). Intuitively, as q increases, γ increases and N^* decreases. The latter has been explained in Example 2(4). For a large q, Model 1 will store many orders in the inventory, accompanied by high inventory maintenance costs. Since the economic improvement γ increases at this point, providing service decomposition for fewer Type 2 customers is more beneficial. When $q > 0.3$, η has the same trend as N^*, which is understandable. Conversely, γ increases with q for $q < 0.3$ implying that $Tc^*(N)$ is more affected by q than C_2.

5. Comparisons of simulation and theoretical results

In this section, we compare the theoretical results with the simulation results to verify the correctness of the analysis. We consider the case of Model 1 and Model 2 for different values of p. Some results of the simulations are given in Figure 9, in which, we assume that $\lambda = 2$, $\mu = 2$, $\alpha = 4$, $\beta = 6$ and $p = 0.7$. We summarize in Table 1 the comparison between the theoretical and simulation values of the performance measures for Model 1 and Model 2. The parameters are set as follows: $\lambda = 10$, $\mu = 10$, $\alpha = 20$ and $\beta = 25$. The notations of the performance measures in Table 1 are described as follows. S denotes the average sojourn time of a stored additional service. I denotes the average number of additional services stored in inventory. W represents the average waiting time of customers. N represents the queue length of customers. We can see in Table 1, the relative differences in the system performances are less than 8%. These numerical results imply the reliability of our proposed model and the derived performance measures.



(a) Simulation of customers' arrival and departure moments in Model 1.

(b) Simulation of the number of customers in Model 1.

(c) Simulation of inventory changes in Model 1.

(d) Simulation of customers' arrival and departure moments in Model 2.

(e) Simulation of customers' waiting time in Model 2.

(f) Simulation of the number of customers in Model 2.

Figure 9. Some simulation results of Model 1 & Model 2.

Table 1. The performance of theoretical and simulation results.

Performance measure			Theoretical result	Simulation result	Relative diff.
Model 1	p=0.2	S	0.6778	0.7191	0.059130933
		I	5.4222	5.3724	0.009226836
		W	0.05	0.0469	0.063983488
		N	1	0.9955	0.004510148
	p=0.4	S	0.4385	0.4186	0.046435655
		I	2.6308	2.4767	0.060342633
		W	0.05	0.0477	0.047082907
		N	1	1.0196	0.019409784
	p=0.6	S	1.2471	1.3405	0.072190447
		I	0.3118	0.3190	0.022828155
		W	0.05	0.0465	0.07253886
		N	1	1.0072	0.007174173
Model 2	p=0.2	W	0.1357	0.1338	0.014100186
		N	2.0071	1.9890	0.009058832
	p=0.4	W	0.35	0.3512	0.003422704
		N	4.3	4.3407	0.009420533
	p=0.6	W	0.35	0.3435	0.018745494
		N	4.3	4.2573	0.009979783

6. Conclusions

In this article, we investigated the efficiency improvement problem of a queueing system with two types of customers through service decomposition. The service of Type 2 customers can be split into two parts. One part is the basic service with the same service rate as the Type 1 customers. When no customers are waiting, additional services are executed after the last customer leaves the system. An additional service can be seen as a pending order stored in the system inventory. The matrix geometric method and the probability generating function were used to perform stationary analysis.

Due to the existence of explicit expressions for the performance measures of Model 1 and Model 2, the impact of service decomposition was also discussed categorically. We demonstrated that the proposed service decomposition is more favorable when Eq (3.17) holds. The idle fraction of the server in Model 1 is shown to be larger than that in Model 2 when the total average time of the split service is less than the average full service time of Type 2 customers. This yields a paradox, where performing additional services during server idle time will increase the fraction of time the server is idle. As for some cases (the total average time of the split service is longer than the average continuous service time of Type 2 customers), the proposed decomposition of services is not beneficial. Therefore, we considered limiting the number of pending orders stored in the system. Similarly, we numerically investigated the improvement of Model 3 versus Models 1&2 in terms of economic lift. The numerical examples indicated that when the total average time of the split service is less than the average continuous service time of Type 2 customers, the total cost function decreases as N increases. This means that managers should provide split services for more Type 2 customers. An intelligent algorithm is used to obtain the optimal inventory capacity and the minimum system cost. The economic improvement achieved by applying the optimal capacity to Model 3 was also graphically presented.

In the future, many interesting directions can be investigated. An interesting direction is that the customer has non-preemption priority to the stored orders. Another direction for future investigation is taking the vacation or working vacation into account when the system is empty. In addition, considering the online shopping model in the context of today's developments, Type 3 customers can be defined as those who place their orders online without basic services.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This research was supported in partial by the Guangxi Natural Science Foundation (No. 2023GXNSFBA026033), the Guangxi Young Teachers Basic Ability Improvement Project (No. 2022KY0191), the National Natural Science Foundation of China (No. 61773014).

Conflict of interest

The authors declare that none of the authors have any competing interests in the manuscript.

References

1. A. Marand, H. Li, A. Thorstenson, Joint inventory control and pricing in a service-inventory system, *Int. J. Prod. Econ.*, **209** (2019), 78–91. http://dx.doi.org/10.1016/j.ijpe.2017.07.008
2. G. Hanukov, T. Avinadav, T. Chernonog, U. Spiegel, U. Yechiali, A queueing system with decomposed service and inventoried preliminary services, *Appl. Math. Model.*, **47** (2017), 276–293. http://dx.doi.org/10.1016/j.apm.2017.03.008
3. G. Hanukov, T. Avinadav, T. Chernonog, U. Yechiali, Performance improvement of a service system via stocking perishable preliminary services, *Eur. J. Oper. Res.*, **274** (2019), 1000–1011. http://dx.doi.org/10.1016/j.ejor.2018.10.027
4. G. Hanukov, Improving efficiency of service systems by performing a part of the service without the customer's presence, *Eur. J. Oper. Res.*, **302** (2022), 606–620. http://dx.doi.org/10.1016/j.ejor.2022.01.045
5. *Wasted time at work costing companies billions in 2006*, Salary.com Staff, 2012. Available form: `http://www.salary.com/wasted-time-at-work-still-costing-companies-billions-in-2006/`.
6. B. Doshi, Queueing systems with vacations-a survey, *Queueing Syst.*, **1** (1986), 29–66. http://dx.doi.org/10.1007/BF01149327

7. J. Ke, The optimal control of an M/G/1 queueing system with server startup and two vacation types, *Appl. Math. Model.*, **27** (2003), 437–450. http://dx.doi.org/10.1016/S0307-904X(03)00047-7

8. Y. Zhang, D. Yue, W. Yue, A queueing-inventory system with random order size policy and server vacations, *Ann. Oper. Res.*, **310** (2022), 595–620. http://dx.doi.org/10.1007/s10479-020-03859-3

9. R. Meena, M. Jain, A. Assad, R. Sethi, D. Garg, Performance and cost comparative analysis for M/G/1 repairable machining system with N-policy vacation, *Math. Comput. Simulat.*, **200** (2022), 315–328. http://dx.doi.org/10.1016/j.matcom.2022.04.012

10. J. Ke, Modified T vacation policy for an M/G/1 queueing system with an unreliable server and startup, *Math. Comput. Model.*, **41** (2005), 1267–1277. http://dx.doi.org/10.1016/j.mcm.2004.08.009

11. P. Vijaya Laxmi, P. Rajesh, T. Kassahun, Analysis of a variant working vacation queue with customer impatience and server breakdowns, *International Journal of Operational Research*, **40** (2021), 437–459. http://dx.doi.org/10.1504/IJOR.2021.114839

12. J. Li, N. Tian, The M/M/1 queue with working vacations and vacation interruptions, *J. Syst. Sci. Syst. Eng.*, **16** (2007), 121–127. http://dx.doi.org/10.1007/s11518-006-5030-6

13. J. Blanc, P. Waal, P. Nain, D. Towsley, Optimal control of admission to a multiserver queue with two arrival streams, *IEEE T. Automat. Contr.*, **37** (1992), 785–797. http://dx.doi.org/10.1109/9.256332

14. A. Turhan, M. Alanyali, D. Starobinski, Optimal admission control in two-class preemptive loss systems, *Oper. Res. Lett.*, **40** (2012), 510–515. http://dx.doi.org/10.1016/j.orl.2012.08.012

15. B. Kim, J. Kim, Waiting time distributions in an M/G/1 retrial queue with two classes of customers, *Ann. Oper. Res.*, **252** (2017), 121–134. http://dx.doi.org/10.1007/s10479-015-1979-1

16. G. Hanukov, A queueing-inventory model with skeptical and trusting customers, *Ann. Oper. Res.*, in press. http://dx.doi.org/10.1007/s10479-022-04936-5

17. G. Lawlor, I'Hôpital's rule for multivariable functions, *The American Mathematical Monthly*, **127** (2020), 717–725. http://dx.doi.org/10.1080/00029890.2020.1793635

18. J. Cohen, The single server queue, In: *North-Holland series in applied mathematics and mechanics*, Amsterdam: Elsevier, 1982.

19. M. Neuts, *Matrix-geometric solutions in stochastic models: an algorithmic approach*, New York: Dover Publications, 1994.

20. G. Latouche, V. Ramaswami, A logarithmic reduction algorithm for quasi-birth-and-death processes, *J. Appl. Probab.*, **30** (1993), 650–674. http://dx.doi.org/10.2307/3214773

AIMS Press